Genes and Proteins Involved in Polysaccharide Colonisation by Marine Microorganisms

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Jennifer Lynne Edwards

September 2009

Table of Contents

List of Figures

List of Tables

Abbreviations

1-D	One dimensional
2-D	Two dimensional
aa	Amino acid
A	Adenine
AnIEX	Anion exchange chromatography
ATP	Adenosine triphosphate
BAC	Bacterial artificial chromosome
BCA	Bicinchoninic acid
BLAST	Basic local alignment search tool
blastp	BLAST search of the protein database using a protein query
blastx	BLAST search of the protein database using a translated nucleotide query
BSA	Bovine serum albumin
C	Cytosine
CatIEX	Cation exchange chromatography
CAZy	Carbohydrate active enzymes
CBM	Carbohydrate binding module
CBD	Cellulose binding domain
CEFAS	Centre for environment, fisheries and aquaculture science
CHAPS	3-[(3-Cholamidopropyl) dimethylammonio]-1-propanesulfonate
CMC	Carboxymethyl cellulose
CTAB	Cetyl trimethylammonium Bromide
ddH_2O	Double distilled water
DOM	Dissolved organic matter
DOC	Dissolved organic carbon
DNA	Deoxyribonucleic acid
dNTP	deoxynucleotide triphoshate
DMSO	Dimethyl sulfoxide

EGC	Ethylene glycol chitin
ESI	Electrospray ionisation
FISH	fluorescent *in situ* hybridization
G	Guanine
GH	Glycosyl hydrolase
HIC	Hydrophobic interaction chromatography
His	Histidine
HPLC	High pressure liquid chromatography
IEF	Isoelectric focusing
IEX	Ion exchange chromatography
IPG	Immobilised protein gradient
IPTG	Isopropyl β-D-1-thiogalactopyranoside
Kan	Kanamycin
LB	Luria Bertani
LC	Liquid Chromatography
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MW	Molecular weight
NCBI	National Centre for Biotechnology Information
OD	Optical density
ORF	Open reading frame
PAGE	Polyacrylamide gel electrophoresis
Pfam	Protein families
POL	Proudman oceanography laboratory
POM	Particulate organic matter
PCR	Polymerase chain reaction
Phusion	Recombinant hi-fidelity DNA polymerase
PI	Isoelectric point

PMF	Peptide mass fingerprint
qPCR	Quantitative polymerase chain reaction
RPC	Reverse phase chromatography
RNA	Ribonucleic acid
mRNA	Messanger ribonucleic acid
rRNA	Ribosomal ribonucleic acid
r.p.m.	Revolutions per minute
SDS	Sodium dodecyl sulphate
SEM	Scanning electron microscope
SIP	Stable isotope probing
SOC	Super optimal broth with glucose
spp.	Species
T	Thymine
TAE	Tris-acetate EDTA
TCA	Trichloroacetic acid
tblastn	BLAST search of the translated nucleotide database using a protein query
tblastx	BLAST search of the translated nucleotide database using a translated nucleotide query
TEMED	Tetramethylethylenediamine
TOF	Time of flight
Tris	Tris(hydroxymethyl) methylamine
UTUCHAPS	Urea thiourea chaps
Q	Quadrupole

Amino Acids

Alanine	A
Arginine	R
Asparagine	N
Aspartic Acid	D
Cysteine	C
Glutamine	Q
Glutamic Acid	E
Glycine	G
Histidine	H
Isoleucine	I
Leucine	L
Lysine	K
Methionine	M
Phenylalanine	F
Proline	P
Serine	S
Threonine	T
Tryptophan	W
Tyrosine	Y
Valine	V

Units

%	Percent
°C	Degrees Centigrade
b	Bases
bp	Base pair
cm	Centimetres
Da	Daltons
g	1 x gravitational force
g	Grams
Gt	Gigaton
h	hours
kb	Kilobases
kDa	Kilodaltons
km	Kilometre
kV	Kilovolts
L	Litre
mg	Milligrams
ml	Millilitres
mm	Millimetre
mM	Millimoles per litre
mV	Millivolts
min	Minutes
MW	Molecular Weight
M	Moles per Litre
ng	Nanograms
nm	Nanometres
OD	Optical Density
pmol	picomoles
r.p.m.	Revolutions per minute

s	seconds
µg	Micrograms
µM	Micromoles per litre
µl	Microlitre
U	one unit of enzyme activity
v/v	Volume/volume
w/v	Weight/Volume
V	Volts
yr	year

Acknowledgements

I would first of all like to thank my supervisors, Professor Alan McCarthy and Professor Clive Edwards to whom I am indebted for their help, support and advice throughout the course of my PhD. I am also extremely grateful to Dr Mark Wilkinson, for all his advice and assistance with proteomic techniques.

Dr Mark Osborne and Dr Ashley Houlden, I wish to thank for their generosity in making available the Colne Estuary Fosmid library. The Proudman Oceanography Laboratory for monthly sampling - thank you for never asking me to go on the boat. I am appreciative to the Natural Environment Research Council (NERC) for providing funding for this work.

I am also grateful to Dr Mike Cox for being an extremely supportive postdoc, particulary in bioinformatics and ofcourse sharing the world of peeling crab shells with me. I would like to extend my thanks to Dr Darren Smith, Dr Paul Fogg, Dr James McDonald and Dr John Kenny for also being helpful and supportive postdocs in the lab and for having patience with me following the many questions I ask on a daily basis. To everyone in Lab H for making work so enjoyable over the last four years, I am really going to miss you all.

I have been fortunate to have great friends in Vic, Claire and Katie who have shown unwavering love and support, you really are the best friends I could wish for. I thank you for the countless nights out we have shared and I apologise for you having to listen to me moan about my PhD over the last four years and my disappearance over the latter months of my thesis writing.

Finally I owe special thanks to my family, particularly my Mum, Dad and brother Carl without whom I would never have completed my thesis. Mum and Dad I cannot thank you enough for your love, support and guidance throughout my life, thank you for your patience, particularly through the latter stages of thesis writing. To Carl I would like to

thank for all his help particularly in preventing me throwing my computer out the window when it would not do what I told it to.

For Nan and Grandad

who never got to see this

Abstract

Polysaccharides are an important source of organic carbon in the marine environment and degradation of marine polysaccharides, such as the insoluble and globally abundant cellulose and chitin, is a major component of the marine carbon cycle. Conversion of these recalcitrant substrates to soluble oligo- and mono-saccharides releases labile carbon from the particulate organic carbon pool, making it available for heterotrophic microorganisms. Although many types of cultured bacteria are known to degrade chitin and cellulose, little is known of the polysaccharide hydrolases expressed by uncultivated chitin and cellulose degrading microbial strains particularly in the marine environment.

Here, molecular biological techniques have been applied to analyse the microbial communities that colonise and degrade insoluble polysaccharides *in situ*. Chitin and cellulose baits were tethered to marine sites in the Irish Sea to act as a matrix for colonising biomass to provide source material for metaproteomic and metagenomic analysis of the biofilm community.

Total DNA was extracted from the cellulose biofilm community and subjected to 454 pyrosequencing. A range of bioinformatics programs were utilised to taxonomically and metabolically characterise the colonising bacteria and the repertoire of glycosyl hydrolases that they possess. A total of 116 sequences were matched to a constructed database of twelve glycosyl hydrolase families.

The biofilm community colonising the cellulose and chitin baits was also analysed for expressed proteins and they were fractionated using 1D SDS-PAGE, 2D SDS-PAGE and chromatographic techniques. Zymograms were employed to visualise cellulase and chitinase activity in biofilm extracted proteins and a 'shotgun proteomics' approach was used for the functional characterisation of a number of peptides.

Bacterial strains were isolated from colonised cellulose baits and screened for cellulase activity; eight isolates were shown to have endoglucanase activity and phylogenetically inferred as belonging to the genera *Glaciecola, Pseudoalteromonas* and *Cellulophaga*. The isolates identifed as *Glaciecola* are particularly interesting as cellulase production has not been previously described in this genus of marine bacteria. Scanning Electron Microscopy (SEM) demonstrated that the biofilm community colonising cotton baits contained an abundance of rod shaped bacteria and *in situ* degradation of the cellulose was observed.

A fosmid library consisting of ca. 7000 clones, constructed from DNA isolated from Colne estuary sediment, was functionally screened using a Carboxymethyl Cellulose (CMC)/Congo red staining method. One fosmid was found to express endoglucanase activity and it was sequenced. Following *in silico* analysis of the ~35 kb fosmid sequence four ORF's of interest were chosen and genes sub cloned. Endoglucanse activity was located on one ORF following screening and the recombinant protein overexpressed predicted to contain a glycosyl hydrolase family 8 (GH8) protein and two carbohydrate family two (CBM2) domains was purified to homogeneity using metal chelation chromatography.

Chapter 1

Introduction

1.1 Microbial Ecology

Microbial communities in the environment perform a plethora of biochemical reactions contributing to primary and secondary production, nutrient transformation and mineralisation of biological compounds which are key roles in the Earth's biochemical cycles and sustaining homeostasis of the ecosystem. The study of these communities, their structure, interactions and function forms the underlying basis of microbial ecology. The majority of our knowledge about microbial ecology has been a result of microbiologists' activities to identify and characterise biochemical processes of microorganisms which can readily be cultured in the laboratory. Microbial communities are however highly complex and dynamic structures and the information obtained from cultured organisms is a poor representation of the genetic and functional diversity present in the environment. Estimates of 1% or less of bacterial diversity is accounted for by cultured representatives in most environments, with a value of 0.001-0.1 % postulated for seawater (Riesenfeld *et al.*, 2004; Amann *et al.*, 1995)

Estimation of bacterial diversity based on isolation of pure cultures is biased, with viable plate counts selecting for certain species and excluding others which cannot be cultured with current methods, skewing diversity and quantification data. Understanding phylogenetic diversity has been improved through the use of the 16S rRNA gene as a molecular marker (for review see Steele & Streit., 2005). Through the 1990's in particular, advances were made in the expansion of our knowledge of the diversity of microbial communities by amplifying 16S rRNA genes using the polymerase chain reaction (PCR) on total DNA extracted directly from many environmental samples. Cloned amplified gene sequences can be subjected to bioinformatic analysis against a database of known

sequences (Amann *et al.*, 1995). This culture independent technique has enabled the existence of many novel species of bacteria to be revealed.

Molecular biological techniques are valuable tools for determining the structure of natural microbial communities. However linking functional diversity with phylogeny has always been a major goal of microbial ecology. Molecular techniques have been used at an accelerating rate over the past decade and the new area of environmental genomics has revolutionised the discipline of microbial ecology. Genomics is being used to firstly obtain information on cultured microbes that play key roles in their environment, as well as on whole community analysis by metagenomics or communitiy genomics, and lately by high throughput analysis using pyrosequencing.

Characterisation of genetic information from organisms in pure culture remains important as it forms the basis of comparative genomics, shaping how we interpret whole community genome sequences and determining the specific roles of groups of organisms in a given environment.

1.2 Techniques in Molecular ecology

Non culture based methods are extremely beneficial for elucidating novel traits, metabolic capabilities and lineages (Karl, 2007). One such branch of molecular ecology is metagenomics, a culture-independent method of directly analysing the metagenomes of microbial assemblages. Because of the known limitations of culture based methods to capture the true genetic diversity of many habitats, techniques have begun to assess the true physiological and metabolic potential of as yet uncultivated microorganisms. This can be achieved by isolating DNA directly from organisms in an environmental sample, cloning large fragments into a vector, and transforming the clones into a cultured host (usually *Escherichia coli)* for further sequence based or functional screening. In order to achieve this, DNA inserts of up to 350 kb have been maintained in BAC (Bacterial artificial chromosome) vectors (Shizuya *et al*, 1992). This approach is a useful tool for studying

gene clusters involved in complete metabolic pathways. This technology is however inherently challenging for applications to microbial ecology, and it is difficult to retrieve large DNA fragments from environmental habitats. Cosmids enable cloning of DNA of 38-50 kb whereby the DNA of interest is ligated into plasmids containing a *cos* site and then packaged into Lambda (λ) phage particles (Collins & Hohn, 1978; Yokobata *et al*, 1991). However, the high levels of instability of genomic inserts in cosmid labraries has led to the use of a low copy number cosmid vector, based on the *E.coli* F factor replicon and known as Fosmid (Kim *et al*, 1992). Fosmid vectors allow for size selection of the cloned DNA fragments of ~40 Kb by packaging the DNA in λ-phage heads, whilst the F factor replicon ensures a stable copy number (Schweder *et al.*, 2008).

DNA sequencing was first developed by Frederick Sanger in 1977, using a chain termination sequencing method. More recently with the introduction of high throughput next generation sequencing technologies such as that of the 454 pyrosequencing platform (Roche) (Margulies *et al.*, 2005) and the corresponding increase in computational capability and analysis (bioinformatics), environmental DNA sequencing has begun to accelerate making the prospect of unlocking true microbial communtiy composition and dynamics realistic.

The next development in the genomics field is sequencing the genome of a single cell (single cell genomics). The technique uses Multiple displacement amplification (MDA) (Dean *et al.*, 2002) in which random oligonucleotide primers initiate DNA synthesis from the original DNA template, and then secondary priming of these products leads to gradual amplification of the genome by a factor of about 10^9 from concentrations of a few femtograms (10^{-15}g) to microgram amounts. This method, previously used in whole genome shotgun sequencing of a cultured *Prochlorococcus* species (Zhang *et al.*, 2006), has the ability to provide whole genome sequences of uncultured bacterial cells, providing that individual cells can be captured from the environment (Dean *et al.*, 2002; Moran, 2008; Hutchinson & Venter, 2006).

Functional genomics can provide supplementary information in microbial ecology. The field focuses on understanding the expression, function, and regulation of genes and proteins and encompasses: environmental transcriptomics, or metatranscriptomics to analyse the collective genes transcribed by a microbial community (Frias-Lopez *et al.*, 2008; Urich *et al.*, 2008); environmental microarrays using genomic and metagenomic DNA used to investigate spatial and temporal patterns of community gene expression and employing probes for phylogenetic or functional evaluation (Park *et al.*, 2008; Zhou , 2002); environmental proteomics or metaproteomics for analysis of the collective proteins expressed in a community, using 1-D and 2-D PAGE along side mass spectrometry to identify expressed proteins (Wilmes & Bond 2004; Ram *et al.*, 2005; Moran 2008).

1.3 The Marine Environment

Approximately 71 % of the Earth's surface is covered by ocean, with an average depth of 4 km, becoming shallow near continental boundaries due to extensive shelves (10-200km) which are a few hundred metres in depth (Suttle, 2007). In comparison to deep ocean water, where physical and chemical parameters are relatively constant, coastal waters are influenced by terrigenous materials (sediments, freshwater, organic carbon and nutrients). The marine environment is therefore a vast, dynamic and complex organisation of habitats which possesses an overwhelming diversity of marine microbial communities and microbial productivity. Habitats range from supercooled brine channels of the arctic ice floes to near boiling hydrothermal vents (Kirchman, 2008). Azam & Malfatti (2007) suggest that with the detection of 10^6 decomposer bacteria per ml seawater, a 1 mm^3 'microenvironment' typically contains all the components of the microbial loop i.e primary producers (cyanobacteria and algae), decomposers (bacteria) and predators (viruses and protists). Collectively bacteria, archaea, protists, phage and viruses interact and contribute to the different microbial ecology roles of the marine environment. Half of the global primary production occurs in the oceans, by the action of phototrophic bacteria and algae which harvest solar energy to produce organic matter

fuelling heterotrophic processes in the ocean (Figure 1.2) (Karl, 2007; Azam & Mefatti, 2007). Heterotrophic bacteria, the most abundant organisms in the ocean are the main contributers to the mineralisation of organic carbon from primary sources, releasing CO_2, nitrogenous compounds and phosphate in the surface layer of sea water (Kirchman, 2008; Azam & Mefatti, 2007). Ninety % of primary production is mineralised in the surface layer of the open ocean with only 10% passing to deeper water, in comparison to coastal areas where more is exported to deeper water or horizontally to less productive waters (Kirchman, 2008).

1.4 Bacteria in the marine environment

Despite the diversity of marine bacteria, the majority are yet to be isolated in culture and there is a paucity of information on true community function and structure (Fuhrman & Hagstrom, 2008). The major known clades are represented in Figure 1.1. Representatives of the *Alphaproteobacteria* such as the genus *Roseobacter* contain cultured as well as non cultured representatives and are a numerically important group which through quantitative 16S rRNA gene clone library analysis have been shown to constitute up to 25 % of the marine bacterial community, and are present in most marine habitats examined (Buchan *et al.*, 2005). The SAR11 clade also represents a diverse group of bacterioplankton, initially described following 16SrRNA gene clone library analysis, and constituting 15% of the total RNA sampled from the Sargasso Sea (Giovannoni *et al.*, 1990). Members of this clade were subsequently isolated through laboratory cultivation by sequential dilution of natural microbial commnuities in very low nutrient (oligotrophic) media (high throughput Dilution-to-Extinction culturing (HTC) method) (Rappe *et al.*, 2002). The method has recently proven successful when 17 new strains of the SAR 11 clade and and the first strain of the SAR116 clade were isolated (Stingl *et al.*, 2007).

The *Gammaproteobacteria* contain one of the best studied groups of marine bacteria- the vibrios, which contain readily culturable species including important polysaccharide-

degraders and pathogens such as *V. cholerae*, and *V. anguillarum* which is an important fish pathogen (Grisez & Ollevier, 1995).

Non-culture based methods have shown that there are still a large number of bacterial clades which are abundant in marine habitats but as yet not represented in culture collections. The SAR86 cluster has yet to be isolated in culture even though their 16S rRNA genes are abundant in 16S rRNA clone libraries (Eilers *et al.*, 2000; Kelly & Christoserdov, 2001). The gene clone cluster, SAR202 is also commonly found in libraries and is both abundant and widely distributed, but not represented in culture collections. Although clone libraries may indicate phylogenetic relationships, functional characteristics can only be implied and while metagenomic and functional genomic studies provide new information on marine microbial ecology, culturing and improvement of culturing methods is still essential. Standard microbiological techniques still have a place in modern microbial ecology and are central to functional and taxonomic assignment as gene sequencing alone is only as informative as the representational coverage of the databases permits.

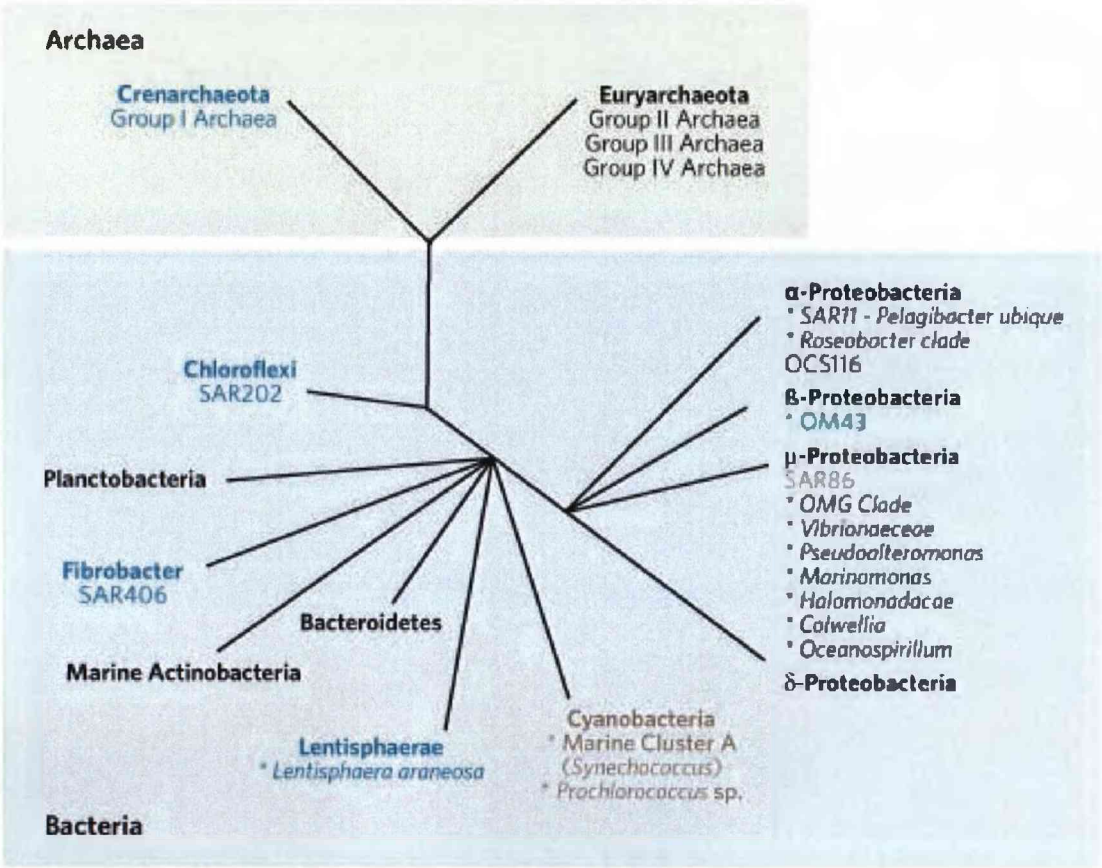


Figure 1.1 Schematic representation of major plankton clades (taken from Giovannoni & Stingl, 2005)

Taxa marked with asterisks represent clades for which cultured members are available. Emboldened text indicates microbial groups that seem to be ubiquitous in seawater. Gold text indicates groups found in the photic zone. Blue text indicates groups confined to the mesopelagic and surface waters during polar winters. Green text indicates microbial groups associated with coastal ocean ecosystems.

1.5 Polysaccharides in the marine environment

It is understood that bacteria are mainly responsible for consuming organic matter in the marine environment, and the carbon cycle depends on this remineralisation of biomass. In marine environments, biomass is mainly produced by phototrophic microorganisms in the upper layer of the water column. Since polysaccharides are major components of biomass, their degradation is essential for the turnover of carbon and nitrogen. No substantial build up of these compounds is apparent, due to continual microbial degradation (Azam & Long, 2001; Miyamoto *et al*, 2002; Yu *et al*, 1991). It is crucial to know the role of microorganisms and their identity in carbon and nitrogen cycling in the marine environment and the rate of degradation of the polymeric substrates.

Polysaccharides, proteins and lipids produced as a result of primary production and by senescence, moulting or excretion from marine organisms together form part of the particulate organic matter (POM) and dissolved organic matter (DOM) pools in the marine environment (Figure 1.2). POM is an important and substantial food source for heterotrophic microorganisms, and aggregates of organic material with variable and dynamic structures are formed. Macroscopic organic aggregates are central to the functioning of marine ecosystems, with aggregation increasing the sedimentation rate of organic matter (Azam & Malfatti, 2007). Aggregates of >0.5mm diameter are known as marine snow in the pelagic environment. This is the main form of large aggregate in the sea, providing environments described as microniches which are enriched in nutrients, trace metals and microbial biomass (Rath, 1998). These are 'hot spots' of microbial activity, colonised by many heterotrophic organisms, metazoans, protozoa, fungi and prokaryotes, particularly bacteria of the *Cytophaga*, *Flavobacteria* and α and γ-Proteobacteria lineages. This distribution of species differs to that of free-living populations in the water column (De Long, 1993; Rath, *et al*, 1998; Simon *et al*, 2002; Smith *et al*, 1992; DeLong, 2007). POM (consisting of plankton and detritus material) and DOM (small particles and colloids) are mineralised and oxidised by heterotrophic marine

bacteria to produce microbial biomass and inorganic products (Figure 1.2) (Nagata, 2008). The global stock of DOM, the most abundant form of organic carbon in the oceans, is thought to comprise 700 Gt (Kirchman, 2008). Heterotrophic microbial metabolism of POM and DOM is therefore central to cycling of carbon in the marine environment. Attached bacteria are part of a microniche and are metabolically more active than free living bacteria, with greater cell specific activity and extracellular enzymatic activities, estimated to be 2.5-10-fold higher in marine snow than in ambient water (Iriberri *et al*, 1987; Long & Azam, 2001; Karner & Herndl, 1992). Attached bacteria also show less seasonal variation in diversity than the free living populations (Iriberri *et al*, 1987).

Particle solubilisation and utilisation of organic carbon and nitrogen are important components of the carbon flux from the ocean surface to its depths. Although hydrolysis of particles is rapid, given the carbon demand of most bacteria, most of the hydrolysate (dissolved organic matter) diffuses away forming plumes that trail behind the sinking aggregates, thus enhancing the growth of free living bacteria (Cho & Azam, 1988; Gram *et al*, 2002 Kiorboe, 2003; Smith *et al*, 1992).

Inter-specific interactions amongst bacteria (e.g. quorum sensing) are known to occur amongst some marine snow bacteria, as well as antagonistic behaviour. It is also suggested that acyl homoserine lactones (AHLs) influence hydrolytic enzyme and antibiotic production, which have been demonstrated *in vitro* (Gram *et al*, 2002).

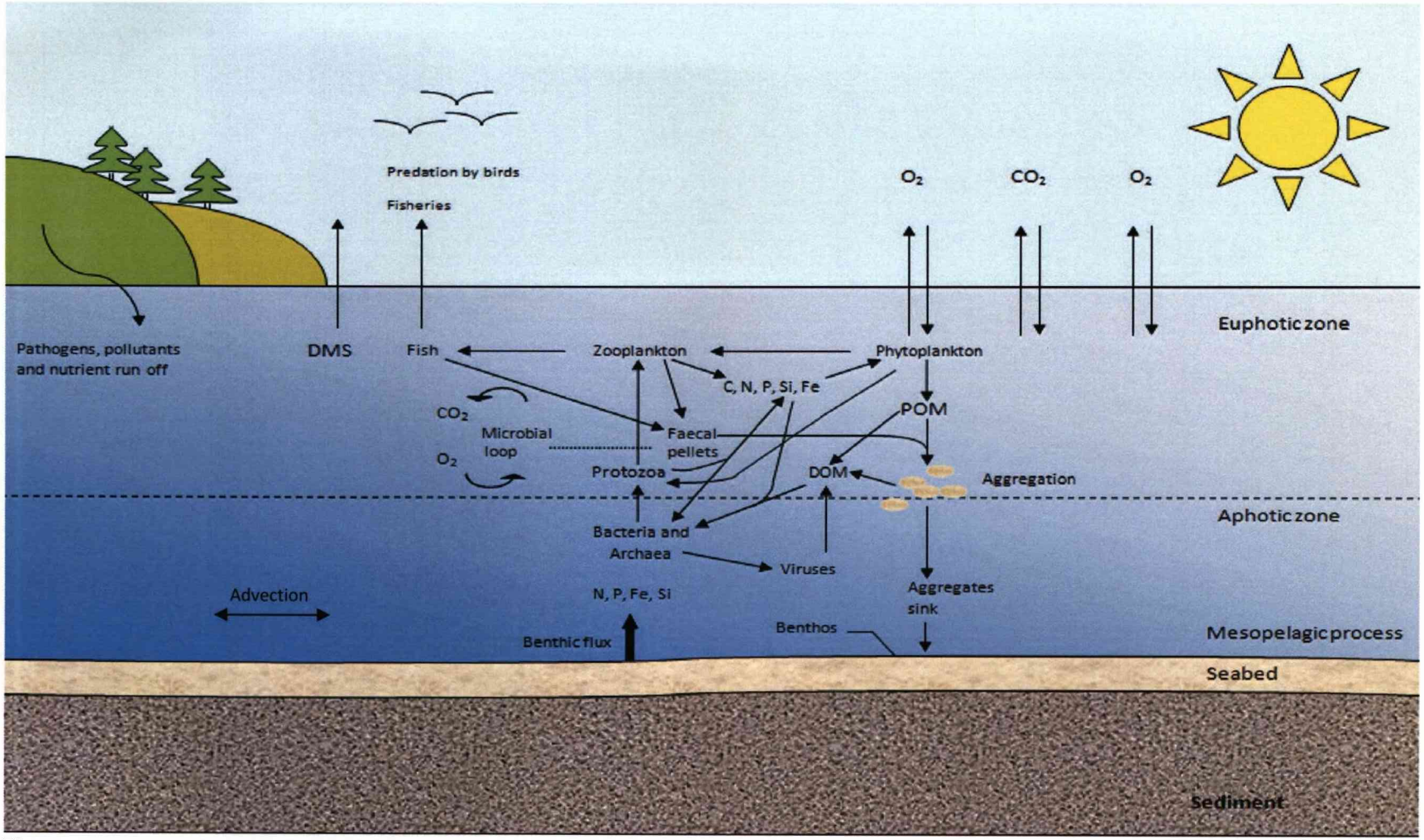


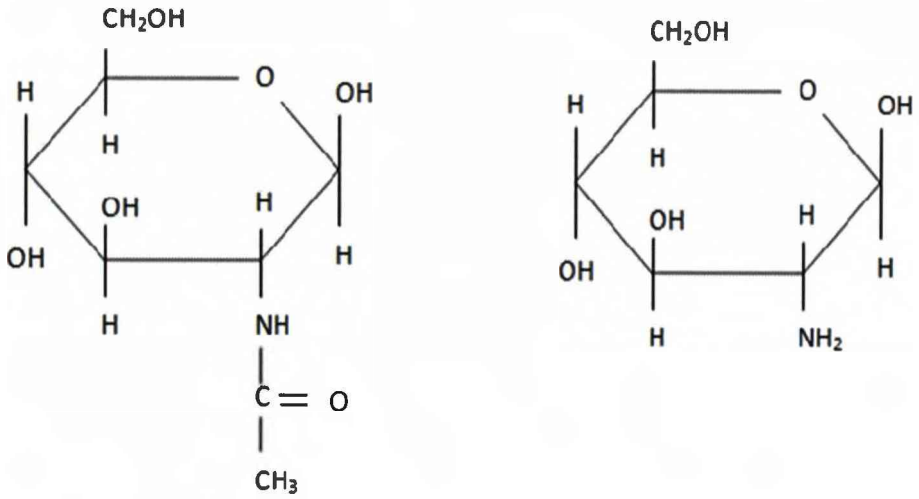
Figure 1.2 Microbial structuring of a marine ecosystem (reproduced from Azam & Malfatti, 2007).

1.6 Chitin

Chitin (Figure 1.4) is a polymer of β-1,4 linked *N*-acetyl-D-glucosamine (GlcNAc)(Figure 1.3) and is the second most abundant polysaccharide on the planet, but the most abundant form of polymeric carbon in the marine environment. It is a structural polysaccharide that is highly cross linked with hydrogen bonds, similar to that of cellulose, differing only by the functional group at position C-2 - hydroxyl (cellulose) and acetamide (chitin). Through hydrogen bonding, chains can orientate themselves differently giving rise to different crystalline structures (α,β and γ) (Figure 1.5), with α-chitin being the most common form in nature where the interacting chains are in an anti-parallel arrangement. In β-chitin, the interacting chains are in a parallel disposition while γ-chitin has a three chain- unit structure in which two chains share the same polarity and the third has the opposite polarity (Martinez & Gozalbo, 2001). Chitin is a primary structural component of exo and endo skeletons and cell wall components, widely produced by protists, fungi, bryozoans, molluscs, annelids, pogonophorans and copepods such as *Tisbe holothuriae* (Kirchner, 1995), diatoms such as *Thalassiosira fluviatilis* (Blackwell *et al*, 1967) and the *Profera* (sponges) (Ehrlich *et al.*, 2007). Availability to bacteria is provided following moulting (exuviae), senescence (carcasses) and faecal pellets such as those from copepods (Yoshikoshi & Ko, 1988). Microbial associations with body surfaces of marine organisms are also quite common (Kirchner, 1995). It is estimated that 10^{10} to 10^{11} tons of chitin are produced annually (Gooday, 1990), with an estimate of 1328 x 10^{6} tons chitin yr^{-1} produced by arthropods in the marine environment alone (Cauchie, 2002) and it is estimated that 10% of marine bacterial biomass production results from chitin degradation (Kirchman & White, 1999).

Unlike other polysaccharides such as cellulose and xylan, chitin contains nitrogen (at typically 5-8%) (Ravikumar, 1999), underlying the importance of recycling of this polymer to return both nitrogen and carbon to the marine environment. Chitin decomposition has been shown to be a rapid and complete process; for example Boyer (1994) reported up to 30 % loss d^{-1} in the York River estuary (USA). Different forms of

chitin, and distinct body parts of the same organism will degrade at different rates depending on the nature of the chitin and levels of associated protein and calcium carbonate (Boyer, 1994). Chitin can also exist in a deacetylated form known as chitosan with varying levels of deacetylation (Figure 1.3), an important factor in the solubility and solution properties of the polymer (Ravikumar, 1999). Chitin and its derivatives have commercial significance for biomedical uses, such as wound healing and medical dressings, pharmaceutical uses in drug delivery, chelating properties and antifungal properties for the agrochemical industry. (For review see Singla & Chawla, 2001; Smelcerovic *et al.*, 2008).



N-acetylglucosamine Glucosamine

Figure 1.3 N-acetylglucosamine and its deacetylated derivative glucosamine.

Chitin is formed by the condensation of multiple N-acetylglucosamine monomers, whilst its deacetylated version forms chitosan

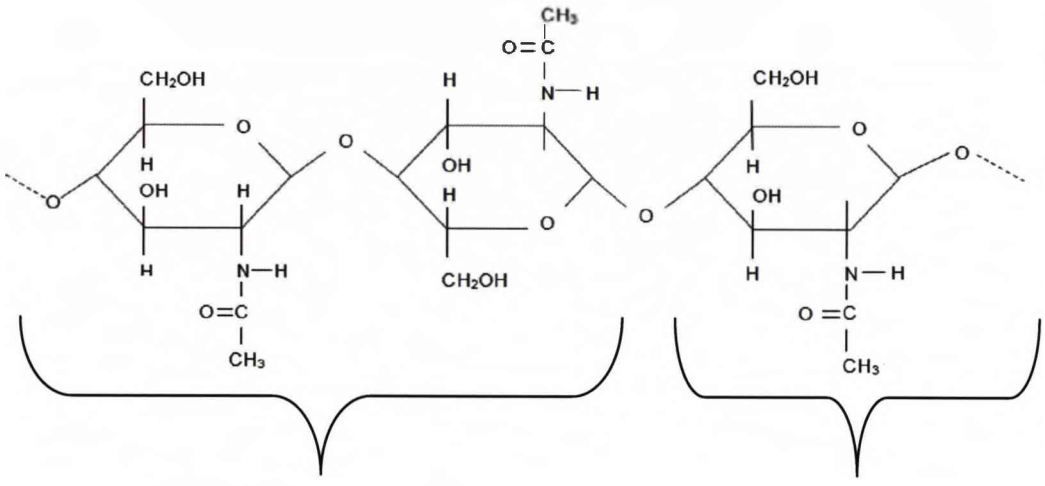


Figure 1.4 The chain structure of chitin.
The N-acetylglucosamine molecules join by condensation reactions forming β-1,4 glycosidic bonds to generate the disaccharide chitobioside, the repeating unit of chitin.

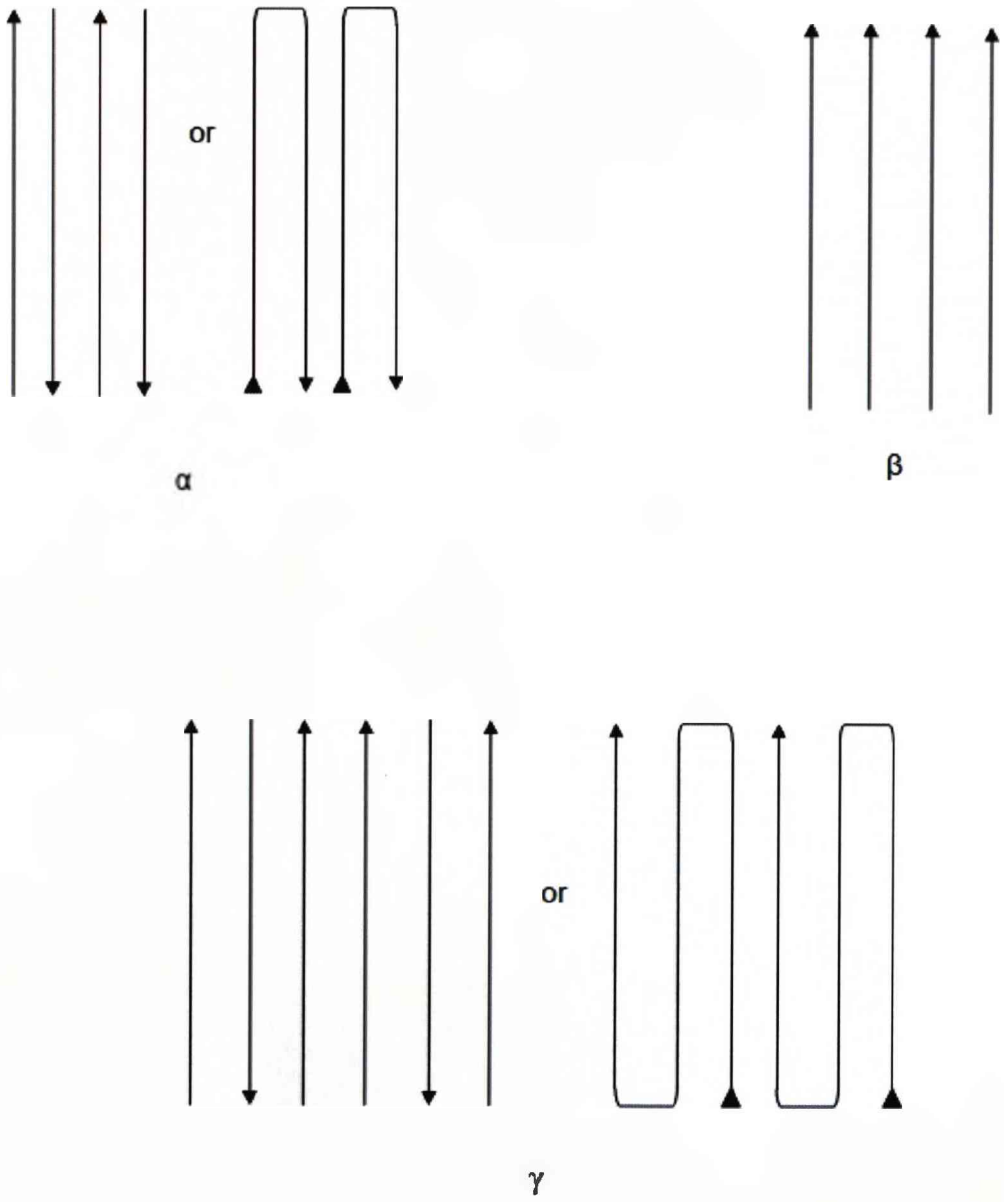


Figure 1.5 Schematic representation of the orientation of the molecular chains in α-, β- and γ-chitin. Taken from Martinez & Gozalbo (2001).

1.7 Cellulose

Cellulose, the main polymeric component of plant biomass, is the most abundant polysaccharide and organic compound on the planet. D-glucose molecules are linked consecutively by β-1,4-glycosidic bonds (Figure 1.6). The multiple hydroxy groups in the molecule stabilise the structure by forming internal hydrogen bonding as well as van der Waals forces (Zhang & Lynd, 2004). Hydrogen bonds are formed between adjacent chains, resulting in largely crystalline aggregates called microfibrils, which in turn combine to form fibrils. Cellulose is a highly crystalline, recalcitrant structure, with a half life under neutral pH in the absence of microorganisms estimated to be ca. 100 million years; chemical hydrolysis requires concentrated acid at high temperatures (Wilson, 2008). The importance of microbial cellulose degradation for the return of carbon to the environment is therefore clearly evident.

Cellulose comprises crystalline and amorphous regions, and the degrees of crystallinity depend on the source (Hilden & Johansson, 2004). Cotton and bacterial cellulose products for example are highly crystalline whilst phosphoric acid swollen and ball-milled cellulose are largely amorphous (Zhang &Lynd, 2004). Cellulose is mainly produced by plants, where it is associated with other polymers such as hemicelluloses, pectins, glycoproteins and lignin. Cellulose is also produced by bacteria such as *Gluconacetobacter xylinus* spp (Kato *et al.*, 2007), algae, fungi such as the oomycetes and animals, for example *Halocynthia* spp. (Aronson & Fuller, 1969; Desvaux, 2005, Schwarz, 2001; Jarvis, 2003). In the marine environment it is the major product of primary production from phytoplankton such as algae and diatoms. The main degraders of cellulose are fungi and bacteria either in microbial assemblages, as individual species or in symbiotic relationships with higher organisms such as termites (Warnecke *et al.*, 2007) or in the digestive tract of ruminants such as cattle and sheep or as symbionts of marine organisms (Yang *et a.,* 2009). Due to the insolubility of crystalline cellulose, cellulolytic organisms secrete enzymes to degrade the substrate with the resulting soluble sugars transported into the cell for further metabolism as a carbon source (Wilson, 2008).

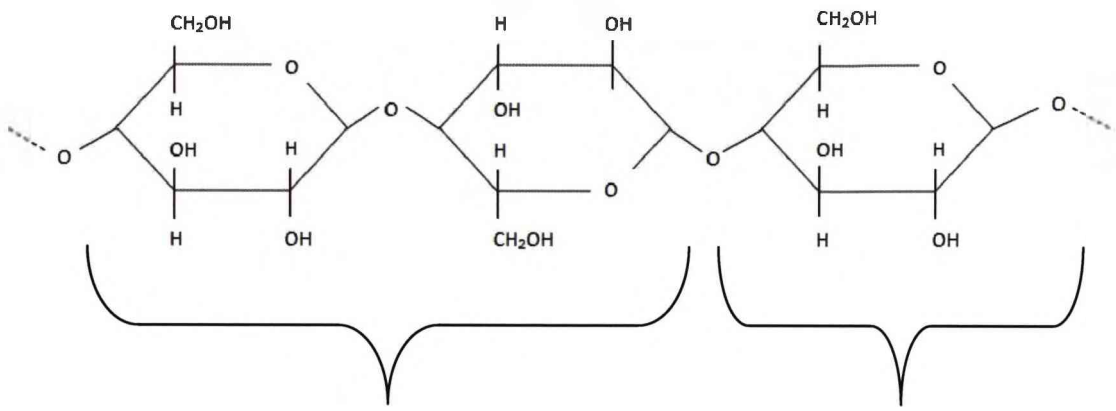


Figure 1.6 Schematic of the chain structure of cellulose.

The glucose molecules join by condensation reactions forming β-1,4 glycosidic bonds generating the disaccharide cellobiose, the repeating unit of cellulose.

1.8 Glycosyl Hydrolases

Cellulose and chitin are both insoluble substrates that require extracellular enzymes, free or cell surface-associated, to convert them to soluble products. Cellulose and chitin are degraded by cellulase(s) and chitinase(s) respectively and these are classified along with other enzymes which hydrolyse the glycosidic bond between two or more carbohydrates or between a carbohydrate and a non-carbohydrate moiety. The standard classification of the enzymes is catalogued at the Carbohydrate Active enzyme web resource (CAZy) (Henrissat, 1991; **(http://www.cazy.org.html)**. Initially classified by the International Union of Biotechnology (1984) based on substrate specifictiy, it has become apparent that enzymes may have various specificities and will hydrolyse the β-1,4 glycosidic bond of more than one type of carbohydrate. The CAZy classification of glycosyl hydrolases (GH's) is based on amino acid sequence and 3-dimensional structure, providing insights into evolutionary relationships and a platform to understand mechanistic properties (Cantarel *et al.*, 2009). In addtion to GH's, the CAZy web resource provides information on all proteins involved in the synthesis and breakdown of complex polysaccharides; for example there are 91 glycosyltransferase families, 19 polysaccharide lyase families and 15 carbohydrate esterase families (Cantarel *et al.*, 2009). To date, 3-dimensional structure representation exists for 34 of the GH families. GH's have a varied architecture, being typically in a modular form composed of one or more catalytic and one or more non catalytic modules, (Ekborg *et al.*, 2007), and can possess a single domain, double domain or multiple domain structure. There are representatives of GH's from *Archaea*, *Bacteria* and *Eukarya* (Henrissat, 1991; **(http://www.cazy.org.html)**. To date, 115 families have been classified on CAZy. These families are further categorised into three topological active site categories; (1) A tunnel demonstrating processive exo-attack; (2) A cleft allowing endo-attack and (3) A crater/pocket suited for degradation of substrates via end-on-attack (Davies & Henrissat, 1995) (Figure 1. 7). Hydrolysis of the glycosidic bond takes place via general acid catalysis requiring a proton donor and a nucleophile/base by one of two modes of enzymatic stereochemistry mechanisms, net

retention or inversion of the anomeric configuration. Retaining enzymes use double displacement mechanisms to catalyse hydrolysis with retention at the anomeric centre, whilst inverting enzymes use single displacement leading to inversion of configuration at the anomeric centre, both requiring the involvement of carboxylic acid pairs (McCarter, 1994; Davies *et al.*, 1997).

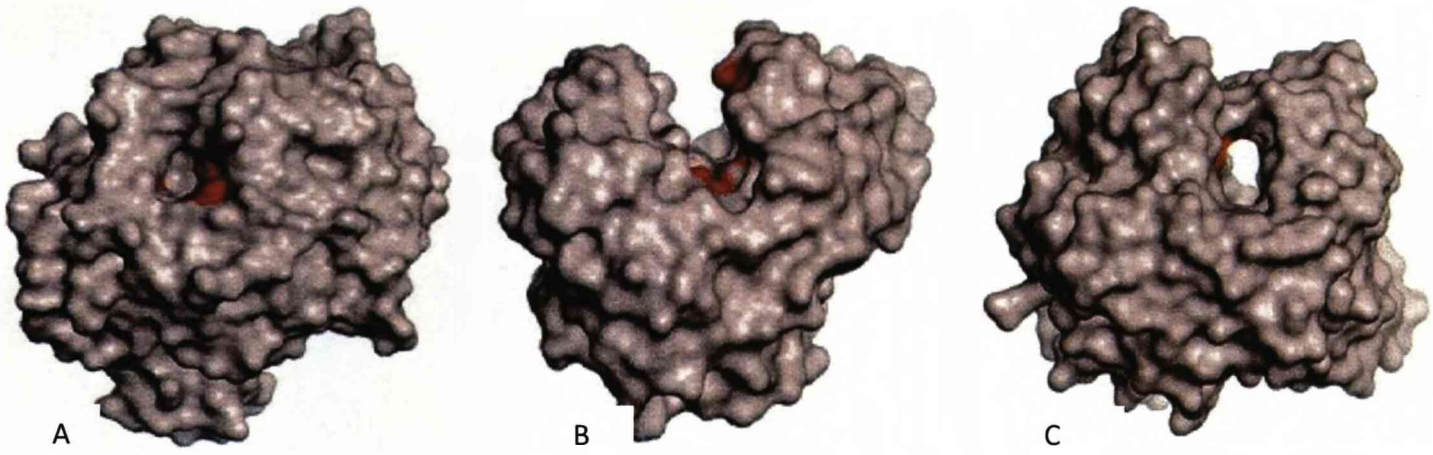


Figure 1.7 Active site topologies of glycosyl hydrolases (from Davies & Henrissat, 1995)

(A) The pocket (glucoamylase from *Aspergillus awamori*) (B) The cleft (endoglucanase E2 from *Thermomonospora fusca*) (C) The tunnel (cellobiohydrolase II taken from *Trichoderma reesei*). Proposed catalytic residues are coloured in red.

1.8.1 Cellulases

Cellulases have representatives currently classified within several of the glycosyl hydrolase families (GH). Cellulose degradation is generally reliant upon the synergistic action of secreted enzymes, whether it be multiple enzymes released from a single species or synergistic action by a consortium of bacteria (Hilden & Johansson, 2004).

Three types of cellulases are generally thought to be involved in complete degradation of cellulose. Endocellulases (Endoglucanase), Exocellulases (Exoglucanase orCellobiohydrolase), and β-glucosidase (Tomme *et al.*, 1996; Sunna *et al.*, 1997; Zhang & Lynd, 2004):

Endoglucanases hydrolyse cellulose fibres randomly in non-crystalline (amorphous) regions, decreasing the degree of polymerisation by attacking interior portions of cellulose molecules. They increase the concentration of chain ends by generating oligosaccharides of varying chain lengths for attack by exo-acting cellobiohydrolases. Primarily they are responsible for chemical changes to the cellulose structure, with a minor role in solubilisation.

Exoglucanases cause erosion of the microfibril surface to reveal further non crystalline regions for endoglucanase attack whilst acting on ends created by endoglucanase in a processive manner. Two classes are known; one attacking the non reducing end of the cellulose chain whilst the second class attacks from the reducing end to produce cellobiose. Therefore they are primarily responsible for solubilisation by generating glucose, or cellobiose as major products, but play a minor role in any chemical properties of residual cellulose.

β-glucosidases catalyse the hydrolysis of soluble oligosaccharides such as cellobiose end products to release the glucose monomer, thus preventing accumulation of cellobiose.

However it is apparent that the mode of action of cellulases is not this simple. It has been shown that endocellulases can also act processively and not just in the random manner first thought. CenC an endoglucanase from *Cellulomonas fimi* appears to act processively from the ends of the cellulose chain following an initial random attack (Tomme *et al.*, 1996), whilst E4 from *Thermomonospora fusca* can bind and cleave internal sites while displaying processive activity cleaving cellotetraose from the non reducing end of the cellulose chain (Irwin *et al.*, 1998).

Even though the basic stucture of cellulose is simple, multiple hydrogen bonding within and between chains produce crystalline and amorphous regions. Cellulose is also rarely found in pure form in nature and is usually complexed with other material depending on the source (Zhang &Lynd, 2004; Wilson, 2008), therefore resulting in a diverse group of enzymes for the degradation of the recalcitrant polymer.

1.8.2 Chitinase

Chitinases are glycosyl hydrolases that convert the polymer (GlcNAc)n to acetate, NH_3, Fructose-6-phosphate and the monomer N-acetylglucosamine (GlcNAc) (Bhattacharya *et al.*, 2007). The ability to metabolise and transport GlcNAc is common as it is the main constituent of the cell wall peptidoglycan in Gram-positive and Gram-negative bacteria (Riemann &Azam, 2002). Chitinases fall into two of the glycosyl hydrolase families (18 & 19) and in bacteria have a size range of ~20-60 kDa (Bhattacharya *et al.*, 2007). Family 18 chitinases have a retaining catalytic mechanism, while family 19 use an inverting mechanism. Due to many forms of the chitin structure (acetylated, deacetylated and varying oligomers) multiple enzymes are required for its degradation; a wide range that have been isolated from a variety of sources. However they can be classsified as belonging to three main groups based on their general activity (Bhattacharya *et al.*, 2007; Cohen-Kupiec & Chet, 1998);

Endochitinase, which cleaves the internal β-1,4-glycosidic bonds of the polymer chain, releasing chitooligosaccharides such as chitotetraose, chitotriose and chitobiose.

Chitobiosidase catalyses the progressive cleavage of chitobioside from the non reducing end of the polymer chain.

β-N-acetylglucosaminidase cleaves oligomeric products of endochitinase and chitobiosidases relasing GlcNAc monomers from the non reducing end of the polymer chain . Both chitobiosidase and β-N-acetylglucosaminidase, are functionally 'exochitinases'.

In addition, chitin deacetylase converts chitin to chitosan, which can then be degraded by chitosanase (Howard *et al.*, 2003). The enzymes involved in the overall degradation, metabolism and transport of chitin are located in the extracellular environment, the periplasm and the cytoplasm. For example the archaeon *Thermococcus chitonophagus* produces a periplasmic chitobiase, a membrane associated endochitinase and an extracellular chitobiase (Andronopoulou & Vorgias, 2004). This is evidence that chitinolytic enzyme production is a tightly controlled cascade system, as observed in members of the genus *Vibrio* (Li & Rosemann, 2004).

Chitinases are ubiquitous in the marine environment, and have varying roles depending on the producing organism, including: defence against pathogens; growth and development; retrieval of nutrients. They have been characterised in a number of marine organisms such as red algae, *Chrindrus verrucosus* (Shirota *et al.*, 2008); vertebrates such as the lamprey *Lampetra japonica*, where the chitinase is thought to play a role in gonodal development and innate immunity (Liu *et al.*, 2009). In the crustacean *Penaeus japonicus,* chitinase is thought to be involved in the molting process (Wilder *et al.*, 1995) and in digestion (Watanabe *et al.*, 1998). Chitinases are prolific amongst marine bacteria, particularly *Vibrio* spp, which is probably the best known culturable group of the

Gammaproteobacteria readily isolated from the marine environment (Kirchman, 2008). This includes *V. proteolyticus* (Itoi *et al.*, 2007; Honda *et al.*, 2008), *V. carchariae* (Pantoom *et al.*, 2008), *V. alginlyticus* (Ohishi *et al.*, 2000) and *V. harveyi* (Svitil *et al*, 1997). Marine archaea (*Pyrococcus kodakaraensis*) (Tanaka *et al*, 1999) and fungi (*Metarhizium anisopliae*) (Kang *et al*, 1998) with chitinolytic activity have also been isolated.

Heterotrophic marine bacteria employ a large number of proteins in the solubilisation and mineralisation of chitin of whichfour main steps are recognised: (1) sensing of chitin either by random collision or by chemotaxis; (2) attachment to the chitin; (3) expression of enzymes; (4) uptake and catabolism of the hydrolysis products (Keyhani & Roseman, 1999).

1.9 Polysaccharide degrading marine bacteria

Cellulose degradation in the marine environment is not well described in the literature. Members of the γ-Proteobacteria appear to be the most notable cellulose degraders in the marine environment. Members of the genus *Pseudoalteromonas* have been cultured and their cellulase(s) identified (Garsoux *et al.*, 2004; Zeng *et al.*, 2006). Most recently two marine bacterial species, *Saccharophagus degradans* and *Teredinibacter turnerae* have been shown to degrade cellulose and their genomes were subsequently sequenced (Weiner *et al.*, 2008; Yang *et al*, 2009). *S. degradans* was isolated from decaying salt marsh grass (*Spartina alterniforna*) from Chesapeake Bay, USA and has become a model organism for marine complex polysaccharide degradation (Ekborg *et al.*, 2005, Taylor *et al.*, 2006). It is known to possess systems to degrade 10 carbohydrates, possessing 13 cellulose depolymerases and 7 accessory enzymes (Taylor, 2006). The annotation of its genome revealed 128 genes involved in the hydrolysis of polysaccharides and 127 genes for carbohydrate binding modules (CBMs). The organism is ranked third out of 400 genomes investigated for the number of GHs encoded (Weiner *et al.*, 2008). *T. Turnerae,* which is found in a symbiotic association with a wood boring mollusc (*Lyrodus pedicellatus*), has also had its complete genome annotated, with 101 GH genes predicted

of which 53 % are thought to be involved in the hydrolysis of wood plant material (cellulose, xylan, mannan, rhamnogalactans) nearly double that of *S. degradans*, together with 117 genes for CBMs. (Yang *etal.*, 2009). Both of these organisms are closely related members of the γ proteobacteria and although *T. turnerae* was isolated from a symbiotic relationship, it is believed that the organism lacks many features of an obligate symbiont (reduced genome size, reduced % GC content, loss of core metabolic genes) and consequently it is thought that *T. turnerae* is a faculative intracellular endosymbiont whose niche also includes, or recently included, a free-living existence (Yang *et al.*, 2009).

1.10 Multi enzyme complexes

It has been shown that in the bacterium *Clostridium thermocellum*, the polysaccharide hydrolase enzymes are part of a 'discrete cellulose-binding multienzyme complex for the degradation of cellulosic substrates' (Lamed *et al.*, 1983). Cellulosome complexes are arranged on the cell surface as polycellulosomal protruberance-like organelles, comprising multiple copies of the cellulosome (Schwarz, 2001). The cellulosome can contain cellulases, hemicellulases, and chitinases amongst other enzymes. Evidence for cellulosomes has to date been found mainly in anaerobic bacteria within the genus *Clostridium* including: *C. thermocellum* (Lamed *et al*, 1983); *C. cellulolyticum* (Gal *et al*, 1997); *C. cellovorans*; *C. acetobutylicum* (Sabathe *et al*, 2002) and *C. Papyrosolvens*. They have been isolated from environments such as soil, wood-chip piles, sewage and the rumen (Doi *et al*, 2003). Cellulosomal components have also been observed in anaerobic fungi such as *Neocallimastix patriciarum* and a *Piromyces* sp. (Ponpium *et al*, 2000; Steenbakkers *et al*, 2002), anaerobic bacteria such as *Bacteroides* spp. (Ponpium *et al*, 2000) and *Ruminococcus albus* (Ohara *et al*, 2000), although the latter is part of the *Clostridium* suprageneric taxon. The complete cellulosome structure has been predicted in a number of species (Bayer *et al.*, 2008). Cellulosome architecture may differ in each organism, however general structures include firstly a major polypeptide called scaffoldin, a subunit of the cellulosome that integrates the other subunits into the

complex, anchored to the cell through an S-layer protein. Attached to the scaffoldin can be a carbohydrate binding module which can directly bind the ultrastructure to substrate or cohesin modules. The latter are functional domains that selectively interact with dockerins, which have a conserved sequence on the catalytic domain. Together, the cohesin-dockerin interaction governs the incorporation of enzymatic subunits into the cellulosomal complex. The cellulosome mediates the interaction between the multiple enzymes into an association with the substrate, minimising diffusion losses of hydrolytic products (Bayer *et al.*, 1998; Shoham *et al.*, 1999; Schwarz, 2001).

Similar functions have been suggested for Gram-negative bacteria. When *S. degradans* is grown with complex polysaccharides as the sole carbon source the development of a surface protruberance resembling cellulosome structures can be observed by scanning electron microscopy (Ekborg *et al.*, 2005). With the recent annotation of the *S.degradans* and *T. turnerae* genomes, a large number of lipoprotein encoding genes have been found, with approximately 180 noted for both genomes (Weiner *et al.*, 2008; Yang *et al.*, 2009). Of these, 23 are thought to be associated with polysaccharide degradation by *T. turnerae* (Yang *et al.*, 2009). Lipoprotein anchors have previously been well studied (d'Enfert *et al.* 1987; Pugsley *et al.*, 1986; Seydel *et al.*, 1999; Weiner *et al.*, 2008). Of the *S.degradans* GHs predicted to be involved in plant cell wall hydrolysis, 32 contain consensus lipobox sequences, of which 5 are cellulases and cellulase accessory enzymes. Additionally on the *S.degradans* genome, two Open Reading Frames have been located with dockerin-like motifs while six ORFs show cohesin-like motifs (Weiner *et al.*, 2008). However most of the information regarding marine cellulose degradation by microorganisms is speculative and much more evidence is required to fully understand the function and diversity of the genes and processes involved.

1.11 Carbohydrate Binding Modules

Non-catalytic modules, initially classified as cellulose binding domains based on the discovery of several modules bound to cellulose (Bolam *et al.*, 1998; Hashimoto, 2006), include proteins isolated from the fungus *Trichoderma reesei* (Tomme *et al.*, 1988) and the bacterium *Cellulomonas fimi* (Gilkes *et al.*, 1988). However a more inclusive name of Carbohydrate Binding Modules (CBM's) was proposed following isolation of modules which bind carbohydrates other than cellulose. The current classification of Carbohydrate Binding Modules refers to a contiguous sequence of ~30-200 amino acids, and they are typically 4-20 KDa within a carbohydrate – active enzyme, and exhibiting a discrete fold with carbohydrate binding activity (http://www.cazy.org/fam/acc_CBM.html; Boraston *et al.*, 2004; Shoseyov *et al.*, 2006). Currently 53 families have been described through amino acid sequence similarity on the CAZy database relating to modules possessing non-catalytic carbohydrate recognition. CBM's have been shown to bind to a wide variety of carbohydrates including cellulose (Simpson *et al.*, 2000), chitin (Itoh *et al.*, 2006) and xylan (Simpson *et al* 1999), with individual enzymes being shown to bind more than one substrate, such as both cellulose and chitin (Ekborg *et al.*, 2007). Primarily the role of CBM's is to bring the catalytic domain into an intimate and prolonged association with the substrate. Thus, by increasing the localised concentration of the catalytic module on the surface of the substrate it increases efficiency and potentiates the ability of the enzyme to provide the catalytic process (Bolam *et al.*, 1998; Boraston *et al.*, 2004; Shoseyov *et al.*, 2006). However, it has also been observed that CBM's may provide additional properties of disruption through non-catalytic means, for example, disrupting the structure of cellulose fibres without any detectable catalytic activity (Din *et al.*, 1991). Arai *et al.*, (2002), showed the requirement of *Clostridium thermocellum* CelJ to have the CBM family 30 for the maintenence of activity of the family 9 cellulase, and for determining the mode of action of some enzymes (processive ability) (Irwin *et al.*, 1998). CBMs are widely distributed and have high ligand variation causing them to be of interest in biotechnology,

with CBM's successfully used as fusion tags as a means of affinity purification of recombinant proteins (Kavoosi *et al.*, 2004; Rodriguez *et al.*, 2004).

1.12 Overall Aim

The identity of the organisms responsible for polysaccharide degradation in the marine environment is largely unknown. The approach described in this thesis is to use *in situ* colonised cellulose and chitin as the source of biological material for metagenomic, metaproteomic and microbial isolation studies directed at organisms, genes and enzymes involved in this primary step in carbon recycling in the Sea.

Chapter 2

Sampling

2.1 Sampling the Irish Sea

The Irish Sea is a semi-enclosed body of water, being part of the Northwest European continental shelf extending northwards to St Davids head and Carnsore Pt on the Welsh and Irish coasts, to the North Channel between Larne and Mull of Galloway. The mean water depth average is about 60m (Dabrowski & Hartnett, 2008). The Liverpool Bay area of the Irish Sea is an area of intense human activity influenced by several rivers (Conwy, Dee, Ribble and the Mersey), causing the area in previous years to be substantially polluted by an expansive chemical industry. Primarily high levels of halogenated compounds have been associated with the area following their use in industrial solvents as well as being biogenically produced (Bravo- Linares, 2007). The Liverpool Bay - Coastal Observatory (Eastern Irish Sea) provides measurements of *in situ* surface waves, vertical profiles of current, temperature, salinity, turbidity, nutrients and chlorophyll. The aim is to understand the response to natural forcing and the consequences of human activity, focussing on the impacts of storms, seasonality and the variations in river discharge (freshwater and nutrients) on the functioning of Liverpool Bay (Bravo- Linares, 2007). The coastal observatory programme uses SMART buoys (CEFAS) maintained by the Proudman Oceanography Laboratory (POL).

Sampling was performed monthly by the Proudman Oceanography Laboratory during maintenace of the CEFAS SMART buoys at two locations in the Liverpool Bay area of the Eastern Irish Sea (Figure 2.1). Baits were provided and used for an *in situ* enrichment for the attachment of polysaccharide degrading biofilms. Cotton string was employed as cellulose bait and crab shell chitin as chitin bait. Both polysaccharide baits were added to custom made nylon mesh bags which were then tethered to the SMART buoys by members of POL (Figure 2.1). Baits were left *in situ* for approximately one month

(Figure 2.1), and following one month the baits were returned to the Microbiology laboratory at the BioSciences building, University of Liverpool. If baits were not used immediately, they were stored at -80 °C.

2.2 Crab shell Chitin pre-treatment

Cleaned crab shells were boiled in 1 M NaOH for 1 h. All NaOH was removed and the crab shells neutralised with dH_2O. The crab shell was covered with1 M HCl, left overnight and again washed with ddH_2O. The surface layer of the crab shell was removed and the crab shell chitin dried in a 60°C oven overnight. Chitin was placed in customised nylon mesh bags (10 cm X 10 cm) for deployment at sites in the Irish Sea (Figure 2.1).

2.3 Cotton string cellulose

Cotton yarn (0.7 cm diameter) (Lancashire cotton best twist- from Texere Yarns, Bradford, UK) was used as cellulose bait. Approximately 1 m of yarn was placed in customised nylon mesh bags (10 cm X 10 cm) for deployment at sites in the Irish Sea (Figure 2.1).

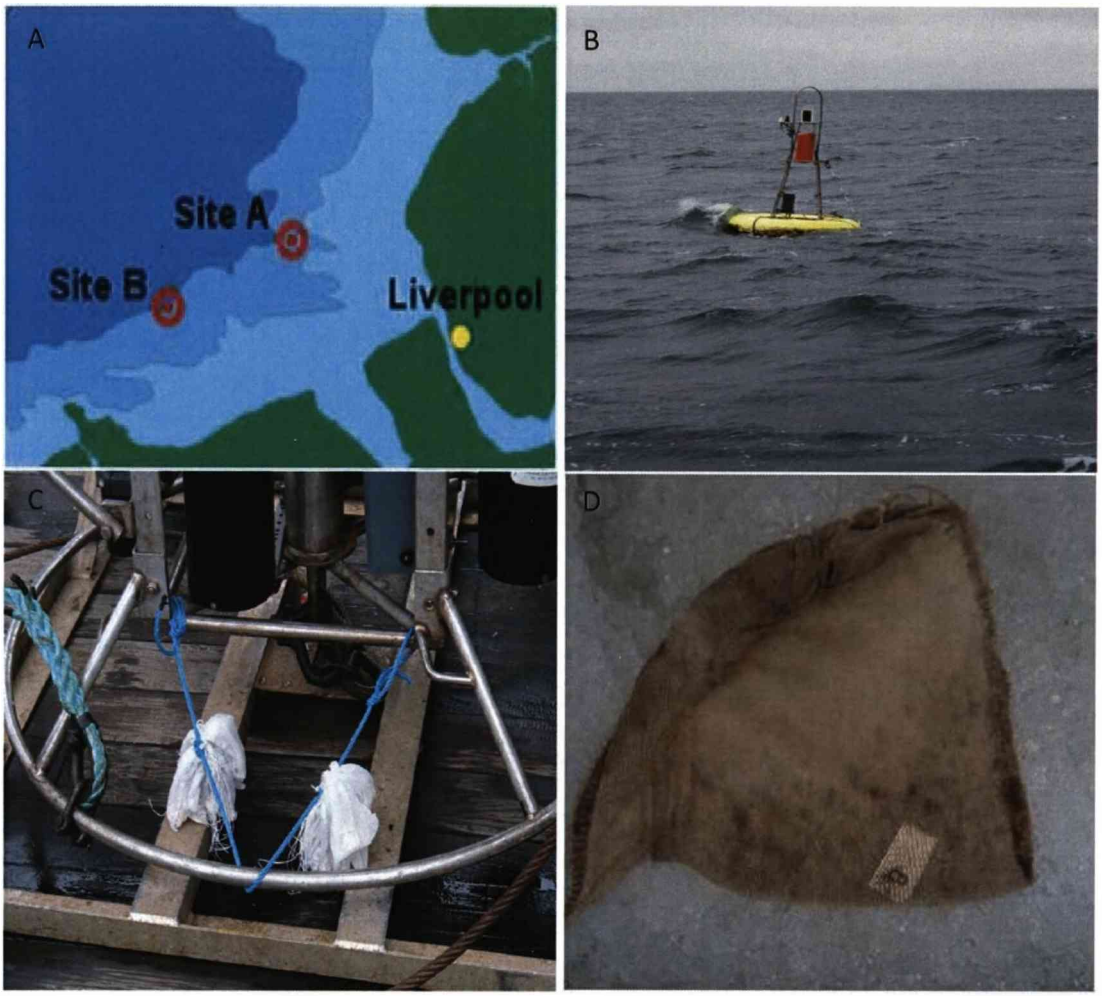


Figure 2.1 Sampling of the Irish Sea

Image showing (A) location of two SMART buoys (site A 53º 32′ N 3º 21.8′ W and site B 53º 27′ N 3º 38.6′ W) located off the coast of Liverpool, eastern Irish sea, maintained by the Proudman Oceanography Laboratory. (B) SMART buoy in the Irish Sea (C) Tethering of the baits to the SMART buoy before deployment into the water. (D) Returned bait bag from the Irish Sea.

Chapter 3

454 Pyrosequencing of the biofilm community colonising cellulose bait in the Irish Sea

3.1 Introduction

3.1.1 Metagenomics

Metagenomics (also known as environmental genomics or community genomics) is described as the analysis of the total genetic material retrieved from a community (Handelsman *et al.*, 1998). Recently the term has come to refer to the application of 'shotgun' sequencing of collective genomes (the metagenome) obtained from an environmental sample, producing randomly sampled sequence data (Krause *et al.*, 2008; Kunin *et al.*, 2008). By direct examination of metagenomic DNA there are two potential benefits. The first is central to microbial ecology; to obtain a comprehensive view of the evolution, lifestyle and diversity of free living microbes, particularly important because the vast majority of microbes are recalcitrant to standard cultivation (Krause *et al.*, 2008). Secondly, metagenomes provide a resource for identifying and exploiting gene sequences for novel enzymes and biologically active molecules for use in the biotechnology industry (for review, see Lorenz & Eck, 2005).

The potential of metagenomics has dramatically increased with the development of sequencing technology. The Sanger sequencing method (1977) developed by Frederick Sanger and Walter Gilbert using a chain-termination method coupled with electrophoretic size separation was used to sequence the first bacterial genome (Fleischmann *et al.*, 1995). Recently a range of next generation, high throughput sequencing platforms have been made commercially available (Rothberg &Leamon 2008; Hall. 2007; Ellegren, 2008). The first of these was the 454 sequencer, developed by Rothberg, Leamon and colleagues

(2008). It was the first commercially available next generation sequencer with the GS-20 released in 2005 by Roche (Margulies *et al.*, 2005) and the first non-Sanger technology used to sequence and assemble a bacterial genome (Margulies *et al.*, 2005). The method has no requirement for the production of a clone library, as used in earlier metagenomic approaches, significantly reducing the cost of sample preparation. Instead, DNA is sheared followed by the attachment of oligonucleotides, in conjunction with a modified pyrosequencing (Ronaghi *et al.*, 1996) method by which nucleotide incorporation is detected by the release of inorganic pyrophosphate leading to generation of photons (Figure 3.1) (Hall, 2007; Rothberg & Leamon., 2008). Initial read-lengths of 100 bp in 2005 have since been superceded by 250 bp in 2007 and in 2008 read lengths reached >400bp (Rothberg & Leamon., 2008). The short read-lengths make pyrosequencing particularly suitable for re-sequencing projects when a scaffold (genome sequence of a relative) is available. The method also reduces bias by the elimination of the need to construct clone libraries prior to sequencing and the reduction in cost compared to Sanger sequencing (Wommack *et al.*, 2008). For example bias may result from products of DNA fragments being cloned may be toxic to the host organism resulting in unsuccessful cloning while using PCR amplified products (i.e. 16S DNA) for library construction restricts the library to the known fraction.

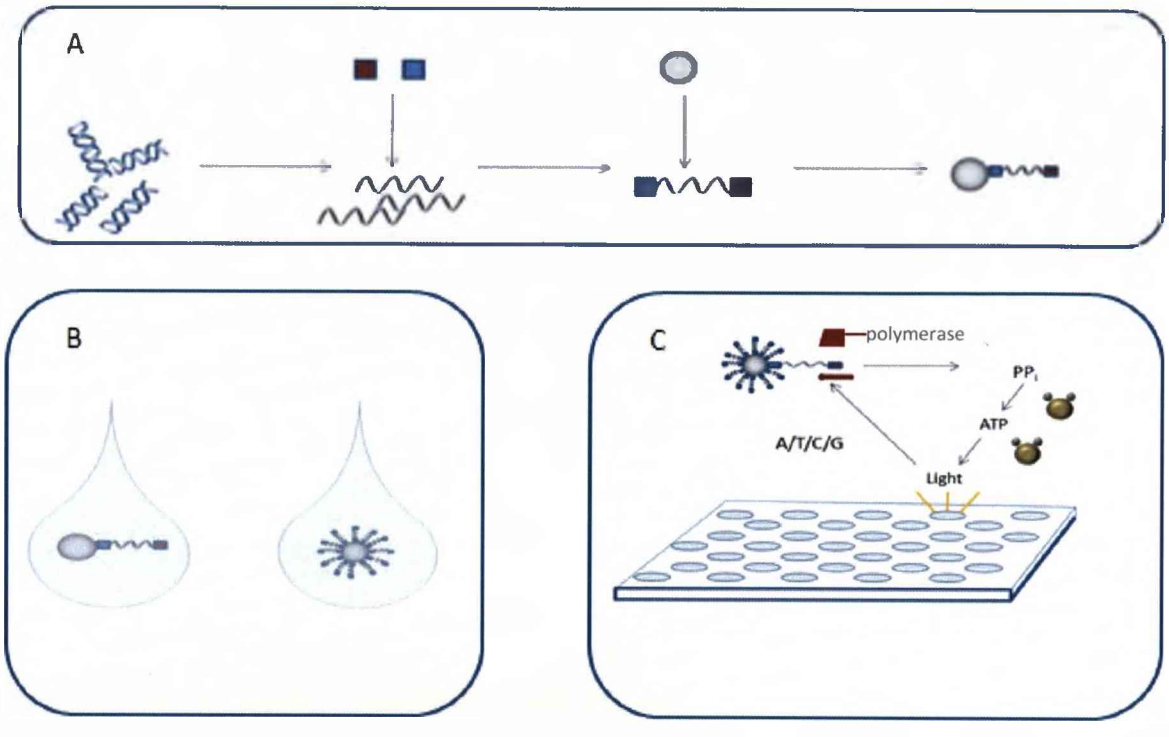


Figure 3.1 Schematic diagram of the 454 pyrosequencing method (adapted from Ellegren, 2008; Medini *et al.*, 2008; Rothberg & Leamon, 2008).

(A) DNA is sheared for use as a starting template and adapters attached specifically to the 3' and the 5' ends. The template fragments are then immobilised to streptavidin-coated beads via a biotin tag on one of the adaptors (one fragment per bead). **(B)** Each bead is then emulsified in a PCR-reaction mixture in oil (one bead per droplet) where fragments are amplified, generating multi copies of the DNA fragments on each bead. **(C)** The emulsion is broken and each bead is deposited into a ~44 µm well of a fibre-optic slide, along with smaller beads carrying immobilised enzymes required for pyrophosphate sequencing. Nucleotide incorporation is detected by the release of inorganic pyrophosphate (PPi) leading to enzymatic generation of photons: PPi is released, converted to ATP and luciferase uses the ATP to generate light, with the process being repeated for each of the four bases. No signal is seen when non complementary nucleotides are added.

3.1.2 MEGAN (Metagenome Analyser)

The computer program MEGAN, is a data management program used in the taxonomic analysis of large sequencing data sets, processing results of sequence comparisons between a known database and metagenome derived sequences. BLAST (blastn, blastx or blastz) comparison output files are analysed by the MEGAN program. FastA headers in the BLAST output corresponding to taxa are identified (according to the NCBI taxonomy). The program uses a Lowest Common Ancestor (LCA) algorithm to assign each read or contig to taxa (Huson *et al.*, 2007). Sequences conserved among a number of species or those susceptible to horizontal gene transfer will be assigned to less specific classification, for example sequences with greater conservation will be placed higher up in the taxonomy (*i.e.* closer to the root) while those more distinct sequences placed on nodes that are more specific *(i.e.* closer to "leaves" representing species). Therefore taxon assignment reflects the level of conservation of the sequence (Huson *et al.*, 2009).

The program has been applied to existing metagenome data by re-analysing data from the Sargasso Sea released by Venter *et al.* (2004) and mammoth DNA (Huson *et al.*, 2007). Using the MEGAN data management program of the 302,692 reads (of mean length 95 bp) 45.4 % were thought to represent mammoth DNA with the remaining reads thought to originate from organisms involved in the putrification process of the carcass. When reads were taxonomically assigned with MEGAN using a bit-score threshold of 30 and atleast two reads assigned 16,972 reads were assigned to bacteria, 761 reads assigned to Archaea and 152 reads assigned to Viruses. Urich *et al.* (2008) used the MEGAN program in the metatranscriptomic analysis of a soil microbial community isolating both rRNA and mRNA for pyrosequencing. From the total 258,411 RNA tags (of ~98 bp) 193,219 had hits against a reference rRNA database which when taxonomically assigned using MEGAN 165,246 were assigned to bacteria and 2,804 to archaea. 21, 133 tags produced hits when compared to the GenBank non-redundant database. An updated version of the MEGAN program which can be used for comparative metagenomics has also been released (Huson *et al.*, 2009), whereby the program allows for the comparison of multiple metagenomic datasets.

3.1.3 MG-RAST

The MG-RAST (Metagenomics-Rapid Annotations using Subsystems Technology) (Meyer *et al.*, 2008) server is based on the RAST server (http://metagenomics.nmpdr.org/), developed by researchers from Argonne national laboratory at the University of Chicago and San Diego State University, for initial use by the National Microbial Pathogen Data resource (NMPDR) community (Aziz *et al.*, 2008). It enables automated annotation of submitted metagenome datasets, providing phylogenetic and metabolic summaries of fragment alignments. MG-RAST, based on the SEED (A genome resource) (Overbeek *et al.*, 2005) uses a subsystems approach to annotation of genomes. RAST was designed specifically for the annotation of completed genome sequences and to provide improved accuracy to high-throughput technology. Subsystems are groups of genes or functional roles consorting in a biological process, for example in a metabolic pathway, which are grouped into a subsystem category, of which there are over 600 (Meyer *et al.*, 2008). Analysis is carried out by comparing sequence data against a number of databases, meaning there is no prior requirement to carry out a comparison search such as one of the BLAST search engines. MG-RAST does not use an output file like other analysis techniques, but the raw data itself or the assembled contigs.

Phylogenetic analysis is carried out firstly whereby sequences of the dataset are compared in all three coding frames in both directions by means of blastx (Altschul *et al.*, 1990) to the SEED non-redundant (nr) database. Protein encoding regions (PEGs) are located and phylogenetic comparisons made based on protein similarities. Alternatively the dataset is compared by means of blastn to a number of rRNA databases including Greengenes (DeSantis *et al.*, 2006) and the RDP II (Cole *et al.*, 2009) databases. Inferences of the dataset can also be made in a metabolic reconstruction analysis. This is performed by functional classification of the PEGs against the SEED FIGfams (protein family's database) and subsystems based on similarity searches (Meyer *et al.*, 2008). MG-RAST contains 299 (May 2009) publically available metagenomes, showing the popularity of the server in metagenome analysis.

3.1.4 AIMS

There are two aims in this chapter; firstly to determine an overall view of the taxonomic diversity of the dataset, using the standalone MEGAN software (Huson *et al.*, 2007) and automated annotation using the MG-RAST (Meyer *et al.*, 2008) server. The second aim, is to identify and determine the diversity of glycosyl hydrolases (GH) within the biological community colonising the cellulose bait recovered from the Irish Sea. This will be achieved by comparison of the 454 pyrosequence dataset with a database of known GH sequences downloaded from the Pfam database and customised for use in this project.

3.2 Methods

3.2.1 DNA extraction (Griffiths *et al.*, 2000)

Nucleic acids were extracted by placing 0.5 g (wet weight) string (retrieved from Liverpool Bay following one month *in situ*, April, 2008) in a Q-biogene purple top multimix tube (lysing matrix E). 0.5 ml hexadecyltrimethylammonium bromide (CTAB) buffer (prepared by mixing equal volumes of 10 % (w/v) CTAB in 0.7M NaCl with 240 mM potassium phosphate buffer, pH8.0) was added along with phenol-chlorofom-isoamyl alcohol (25:24:1; pH 8.0). Cells were lysed by bead beating in a Ribolyser for 30 s at a speed of 5.5 m/s, and the aqueous phase containing nucleic acids separated by centrifugation (16,000 × g) for 5 min at 4°C. The aqueous phase was transferred to a fresh microfuge tube and phenol removed by mixing an equal volume of chloroform-isoamyl alcohol (24:1), followed by centrifugation at (16, 000 × g) for 5 min. Nucleic acids were obtained by precipitation of the top layer by the addition of 2 volumes of 30 % polyethylene glycol (PEG) solution (30 % polyethylene glycol & 1.6 M NaCl), incubated overnight at 4 °C. The precipitated nucleic acids were collected by centrifugation (16, 000 × g) for 15 min. The supernatant was removed and the pellet washed with 200 µl 70 % ice cold ethanol and air dried prior to resuspension in 50 µl sterile ddH_2O

3.2.2 454 Sequencing

DNA was sequenced using the 454 Corporations GS-FLX instrument at the NERC-funded Advanced Genomics Facility at the University of Liverpool (http:/www.liv.ac.uk/agf/) by the method of Margulies *et al.*, 2005.

3.2.3 BLASTX

The Irish Sea cellulose biofilm DNA 454 pyrosequencing dataset contained 223,263 reads ranging from 8 bp to 375 bp in length, providing a total of 48,338,140 bp of DNA. The raw randomly pyrosequenced reads were assembled into 26,860 contigs by the bioinformaticians at the Advanced Genomics Facility at the University of Liverpool and provided in FASTA format. All contigs were translated in all three reading

frames in both directions and compared against a downloaded version of the NCBI Non-Redundant (nr) protein database. The NCBI-nr protein sequence database contains entries from GenPept, Swissprot, PIR, PDF, PDB and RefSeq excluding environmental sequences. It is non-redundant in the sense that identical sequences are merged into a single entry (downloaded January 2009). An 'all against all' BLAST comparison was performed via blastx (Altschul *et al.*, 1990), using the command (below) recommended by the bioinformaticians at the Advanced Genomics Facility at the University of Liverpool and implemented with Bio Linux 4 (http://nebc.nox.ac.uk/tools/bio-linux/bio-linux-5.0). Blastx was run from the command line for the 454 sequences using the NCBI-nr database with a word hit extension threshold of 999 (option –f, is used to set a compromise between speed and sensitivity. Requiring there to be greater alignment before it assigns it as a hit (this is less sensitive but quicker)), a multiple hits window size of 4 (option –A, is a setting for the number of processors used (speeding up the computer)), gap opening cost of 32767 (option –G) and a gap extension cost of 32767 (option –E). –G & -E commands the use of the BLOSUM 62 matrix, not allowing for gaps in the alignment to increase the speed of the process the parameters are set high due to the large number of short reads contained in the dataset.

Blastall –p blastx –i (454 sequences) **–d** (nr database) **–o** (output file) **–f999 –A 4 –G 32767 –E32767**.

3.2.4 MEGAN

The output file of blast alignments resulting from the blast 'all against all' of Irish Sea DNA 454 sequence data and the NCBI-nr database was loaded into a windows version of the MEGAN program version 3.2.1 and the LCA algorithm applied to compute the assignment of contigs to taxa. The LCA parameters used were: min support: 5; min bit score: 35.0; top percent 10.0; win score 0.0 (For LCA algorithm see 3.1.2).

3.2.5 MG-RAST

All cellulose bait DNA 454 assembled contigs were uploaded in a FastA format to the MG-RAST server at the SEED (http://metagenomics.theseed.org; Meyer *et al.*, 2008) on 20/02/09. The dataset was submitted under the name LpoolBay454 and was assigned the Metagenome ID:4442594.3.

3.2.6 Glycosyl Hydrolase Database construction

The glycosyl hydrolases (GH) possessing representatives of cellulase (endoglucanase) and chitinase were identifed by analysing the GH families in the Carbohydrate Active Enzyme (CAZy) web resource. The protein sequences of GH families 5, 6, 7, 8, 9, 12, 16, 18, 19, 45, 48 and 61 were downloaded (February 2009) from the Pfam database (http://pfam.sanger.ac.uk/) (Table 3.1). All metagenome derived contigs were used as a query and compared against the GH database with blastx using the same parameters as described 3.2.3.

Table 3.1 Glycosyl Hydrolase families downloaded from Pfam

Glycosyl Hydrolase Family (CAZy Family)	Pfam family ID	Number of sequences
5	PF00150	1210
6	PF01341	150
7	PF00840	253
8	PF01270	199
9	PF00759	633
12	PF01670	137
16	PF00722	1053
18	PF00704	2592
19	PF00182	731
45	PF02015	101
48	PF02011	46
61	PF03443	183

Sequences were downloaded from Pfam (February 2009) for the construction of a glycosyl hydrolase database to compare against contigs generated by the 454 sequencing of the Irish Sea cellulose bait biofilm

3.3 Results

3.3.1 454 Sequence output

The 454 pyrosequencing returned 223,263 reads of DNA sequence, containing 48,338,140 bp of DNA with a size range of 8-375 bp. The raw read data was assembled by the bioinformaticians at the advanced genomics facility, University of Liverpool into 26,860 contiguous sequences (contigs) with a size range of 93-26,859 bp consisting of 6,841,343 bp. Of the assembled contigs in the dataset, the majority were less than 1 kb in length, suggesting significant heterogeneity within the sample. Therefore a 'gene centric' approach to the data analysis was taken, which was largely based on blastx annotation limiting the dataset to the 'known fraction' (Raes, 2007). BLAST sequence analysis to identify homologs of queries against reference sequences can be computationally intensive especially with metagenome datasets. Blastx was chosen from the group of Basic Local Alignment Search Tools because it identifies similarity in a translated query to a protein database. This was necessary because the reading frame of environmental contigs is unknown, and comparison at the protein level is more sensitive. Although computationally more expensive than other search tools such as blastn which compares a query DNA sequence against a reference DNA database. It is less computationally intensive than the tblastx tool which locates similarities between a translated nucleotide database and a translated nucleotide query. The blastx parameters (recommended by the bioinformatics support at the Advanced Genomics facility, University of Liverpool) of the all-against-all search presented here are set as a compromise between sensitivity and speed. Even with the compromised parameters (3.2.3), the blastx search process took over four weeks. Clearly increasing sensitivity of the search at this stage would have required a period of computation time and resource that was not available.

3.3.2 MEGAN

All assembled contigs (26,860) were translated in all six reading frames and compared to a downloaded version of the NCBI nr protein database (downloaded January 2009), using blastx. In accordance with MEGAN recommendations (Huson *et*

al., 2007) relaxed alignment parameters were used for the blastx search. Of the 26,860 contigs, 21,538 provided hits with the blastx parameters set. These results were loaded into the MEGAN program version 3.2.1 (Huson *et al.*, 2007) and the LCA algorithm applied to compute an assignment of contigs to taxa, leading to an estimation of the taxonomic distribution of the contig sequences. The LCA algorithm assigned 19,810 contigs to taxa, 1,610 remained unassigned (matched sequences in the nr database but did not meet the threshold of the LCA algorithm). This feature of the program reduces the chance of obtaining false positives (Huson *et al.*, 2007). 5,322 contigs had no hits when compared against the NCBI nr database which is probably a result of the limited sequence diversity present in reference databases compared to the repertoir of genes in the natural environment.

17,309 contigs (64 %) of all contigs were assigned to the Bacteria, which were therefore predominant in the metagenome. When taking only those contigs that could be taxonomically assigned 87 % were assigned to bacteria. 979 contigs were assigned to eukaryota. No contigs were assigned to Archaea or viruses (fig 3.2).

Of the 17,309 contigs taxonomically assigned to Bacteria, 2731 could not be resolved to a lower taxon within the domain. This can arise when high levels of conservation are present in sequences (Huson *et al.*, 2007). Although 10 phyla of bacteria are represented in the dataset, their relative predominance varies. Numerically dominant phyla by far are the Proteobacteria (9,470) and the *Bacteroidetes* (4,682). Therefore of the contigs assigned to the bacteria 55 % were assigned to the Proteobacteria and 27 % to the *Bacteroidetes* with these two phyla combined accounting for 82 % of all contigs assigned to bacteria (Figure 3.3). Planctomycetes (37), Spirochaetes (13), Firmicutes (73) chloroflexi (10) Cyanobacteria (46), Actinobacteria (16) Fibrobacteres/Acidobacteria (16) and the Chlamydiae/Verrucomicrobia group (141) are all represented albeit at lower numbers (figure 3.3). (The numbers in brackets represent the number of contigs assigned per taxon).

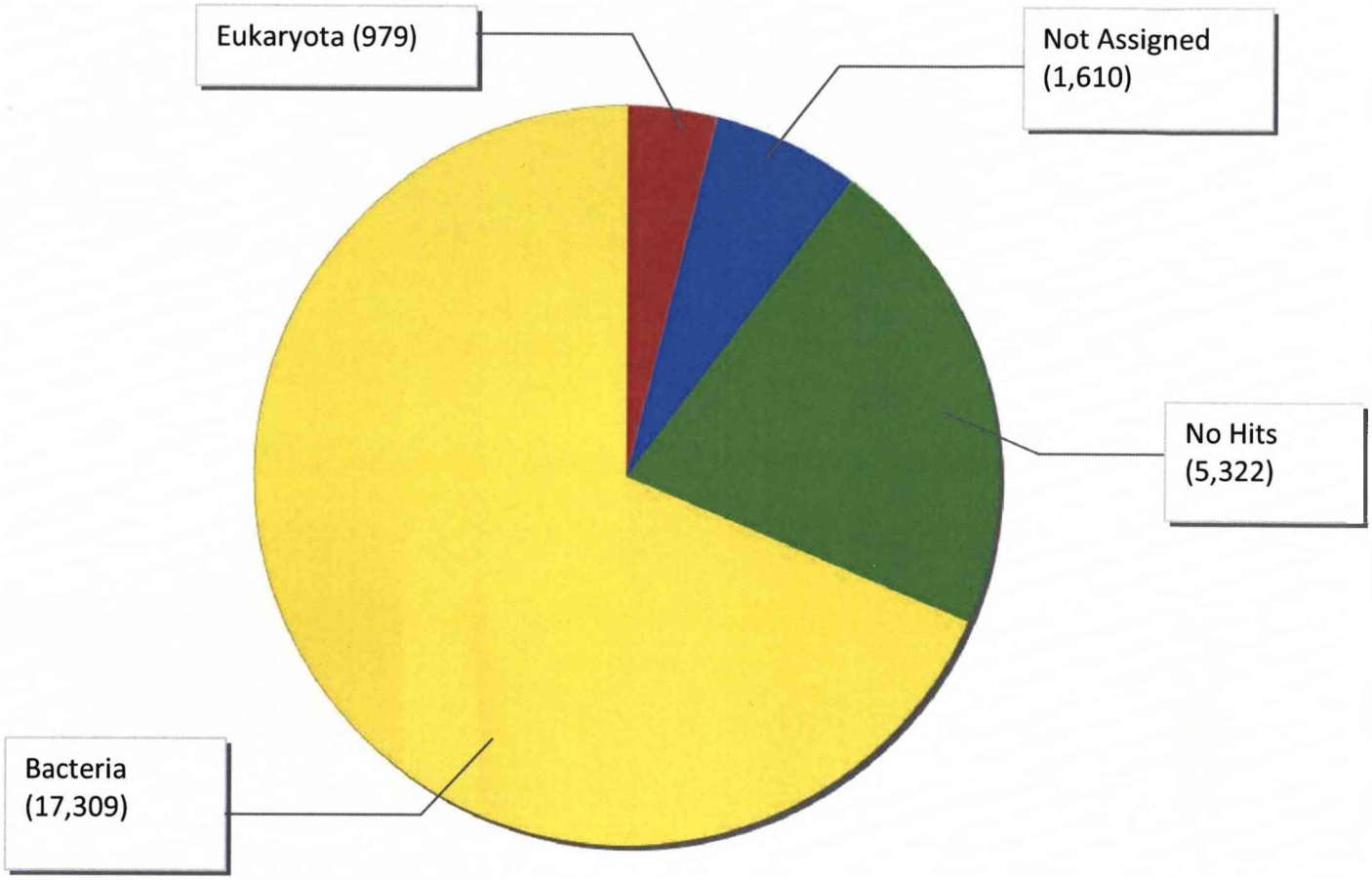


Figure 3.2 Taxonomic assignment of contigs at the Domain level

All contigs (26,860) were taxonomically assigned based on the blastx search of Irish Sea 454 pyrosequencing derived concontigs against the NCBI-nr database.

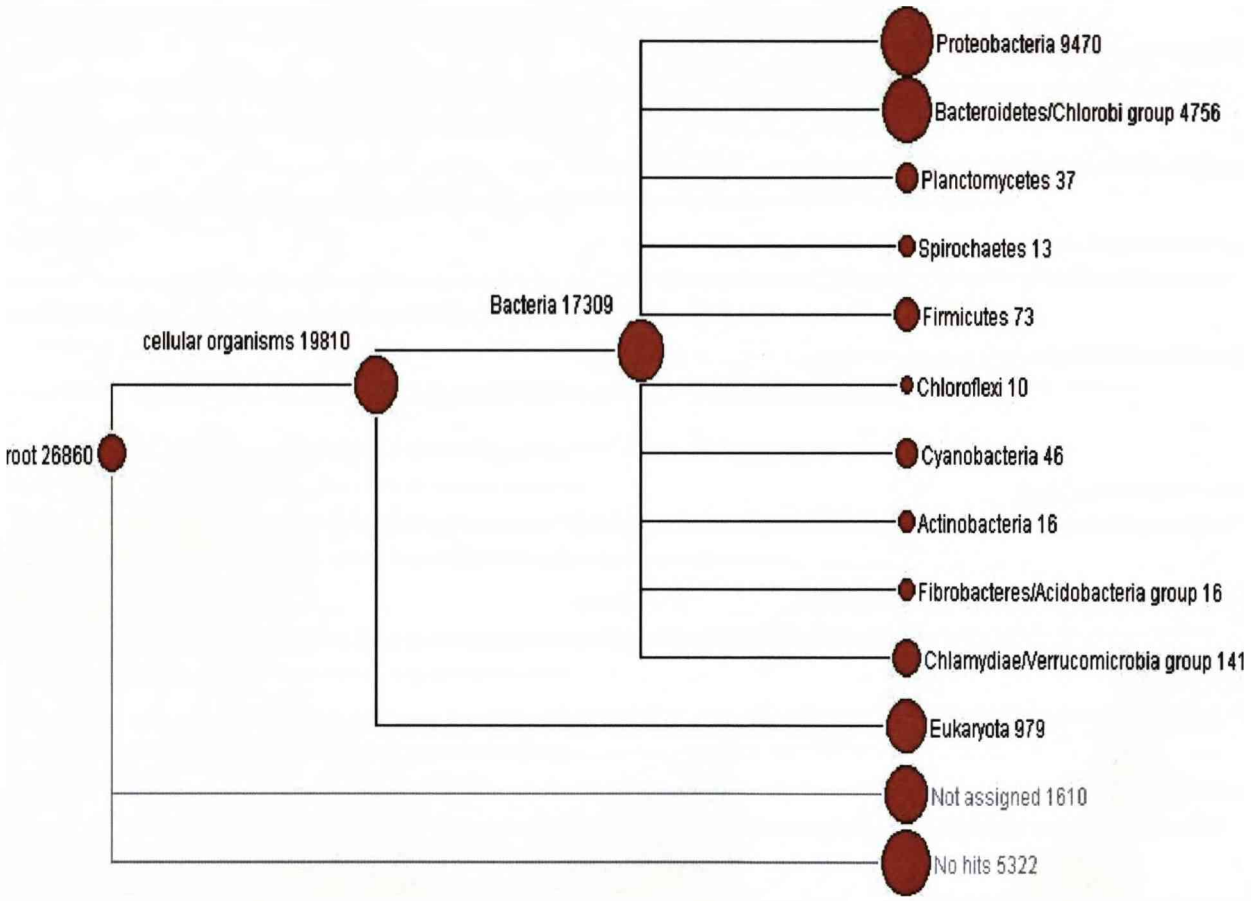


Figure 3.3 Taxonomic tree of the bacterial phyla found in the Irish Sea cellulose bait DNA 454 assembled contigs

Analysis of 26,860 assembled contigs based on blastx comparison against the NCBI nr database and taxonomically classified by MEGAN. The numbers of contigs assigned to each taxon is cited.

When resolving analysis further of the Proteobacteria, *Gammaproteobacteria* contained the most frequent assignment of contigs (6284), accounting for 36 % of all those assigned to the Bacteria (figure 3.4). The *Alphaproteobacteria* were also highly represented in the dataset (1380). Of the *Bacteroidetes*, the *Flavobacteria* (1747) and the *Sphingobacteria* (1177) predominated (figure 3.4). However, 1535 contigs assigned to the Proteobacteria and 1539 contigs to the *Bacteroidetes* could not be assigned at a level lower (Figure 3.4).

Figure 3.5 represents the distribution of contigs of the two most dominant classes within the sample of the Proteobacteria. The majority of contigs assigned to the *Gammaproteobacteria* (45 %) were assigned to the *Alteromonidales* (2856) of which 49 % were accounted for by assignment to *Pseudoalteromonas atlantica* (*Pseudoalteromonas*) and *Saccharophagus degradans* (*Alteromonadaceae*). Assignment of contigs to the *Alphaproteobacteria* was dominated (78 %) by the *Rhodobacterales* (1080), of which 44 % of contigs could not be assigned a further taxon (Figure 3.5).

Of the contigs assigned to the *Flavobacteria* (1747) the majority could be further assigned to the *Flavobacteriales* (1609/92 %) (figure 3.5), whilst 710 (60 %) of the 1177 contigs classified within the *Sphingobacteria* could be attributed to members of the *Flexibacteraceae* (Figure 3.6), of which 36 % and 17 % of *Sphingobacteria* matches were attributed to *Cytophaga hutchinsonii* and *Microscilla marina* respectively (Figure 3.6).

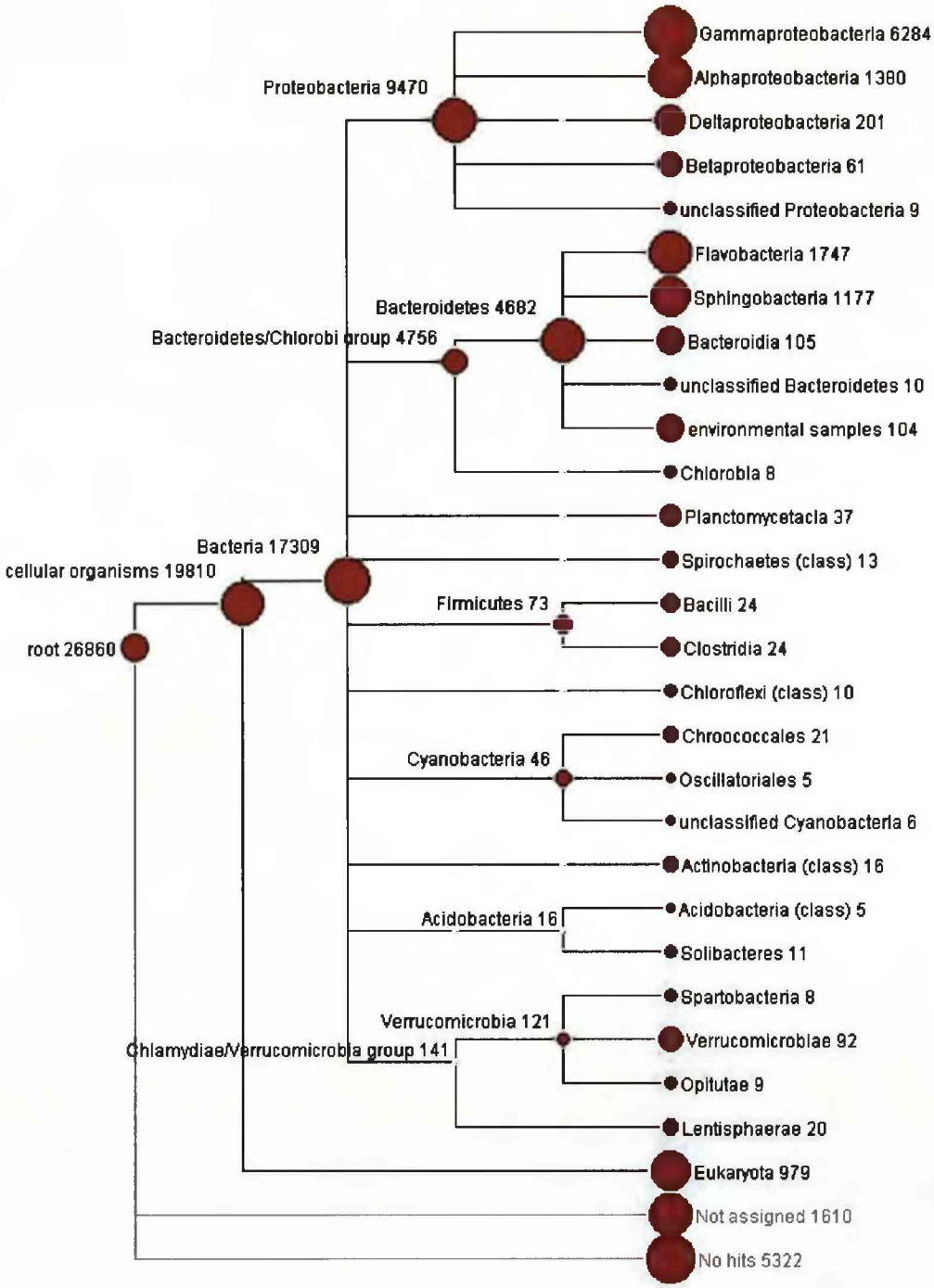


Figure 3.4 Taxonomic tree of the bacterial Classes found in the Irish Sea cellulose bait DNA 454 assembled contigs

Analysis of 26,860 assembled contigs based on blastx comparison against the NCBI nr database and taxonomically classified by MEGAN. The number of contigs assigned to each taxon is cited.

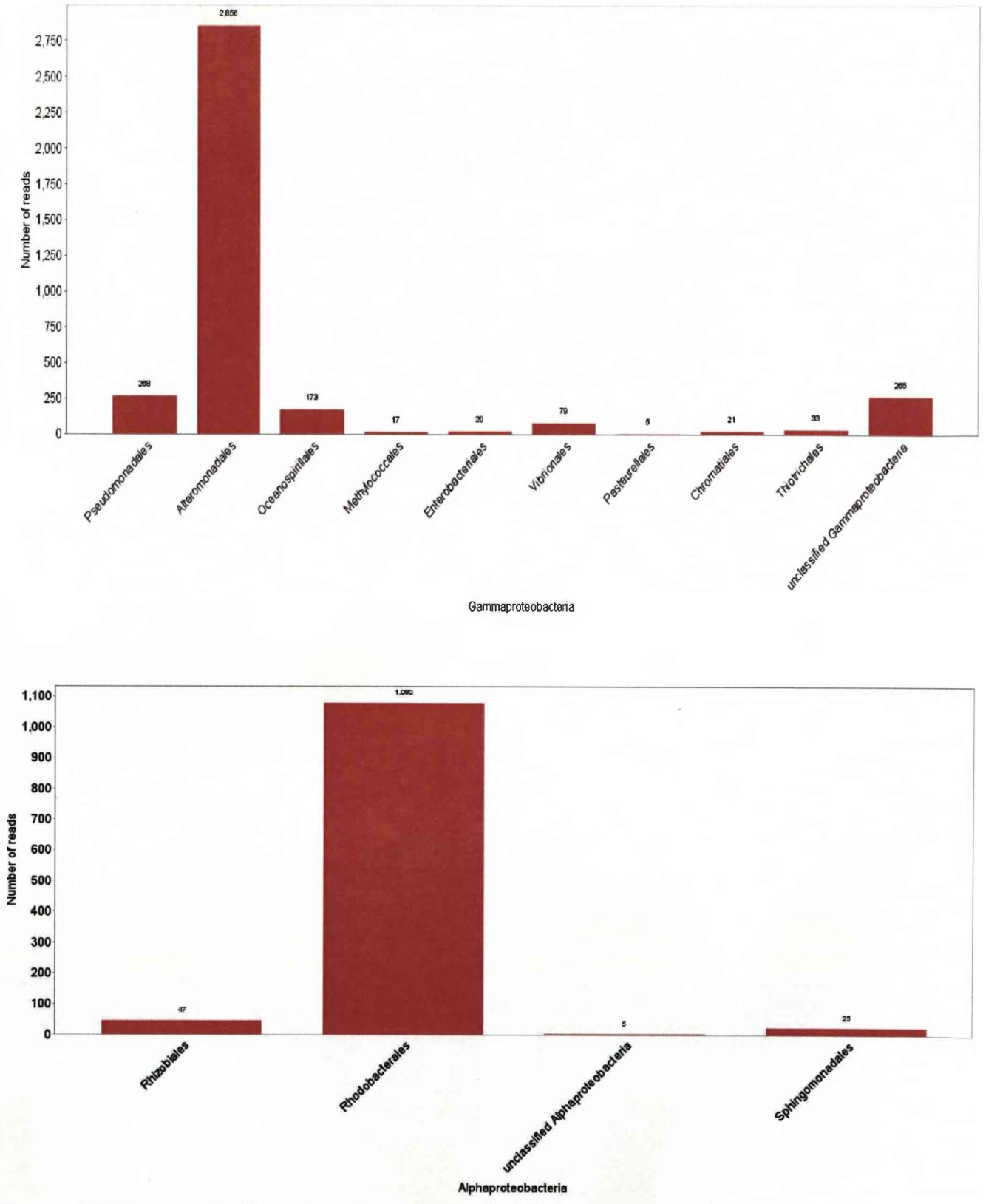
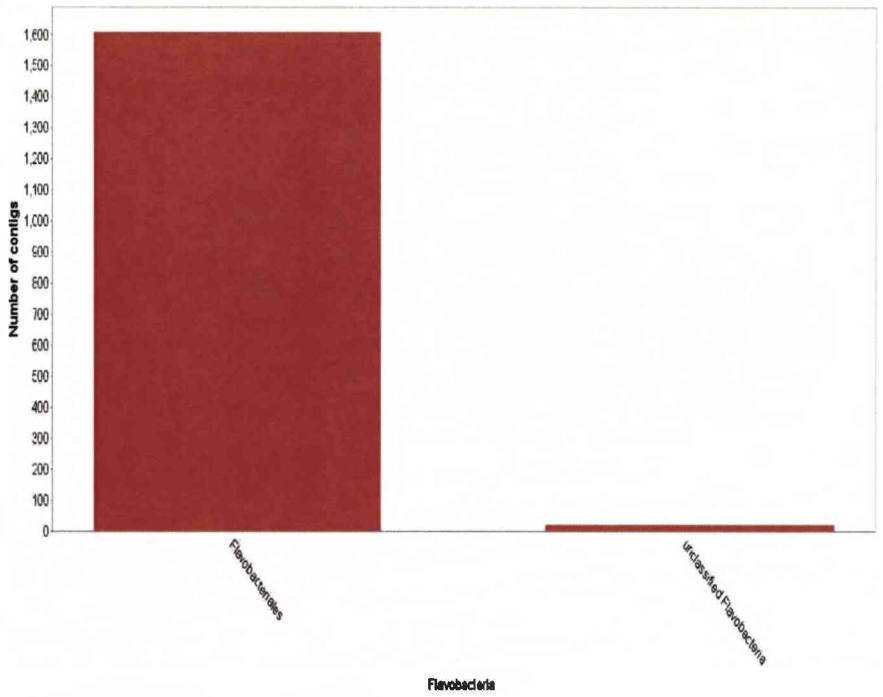


Figure 3.5 Taxonomic diversity of contigs at the Phylum of proteobacteria

Contigs (indicated as number of reads) distribution within the classes of (A) *Gammaproteobacteria* and (B) *Alphaproteobacteria* as computed by MEGAN.



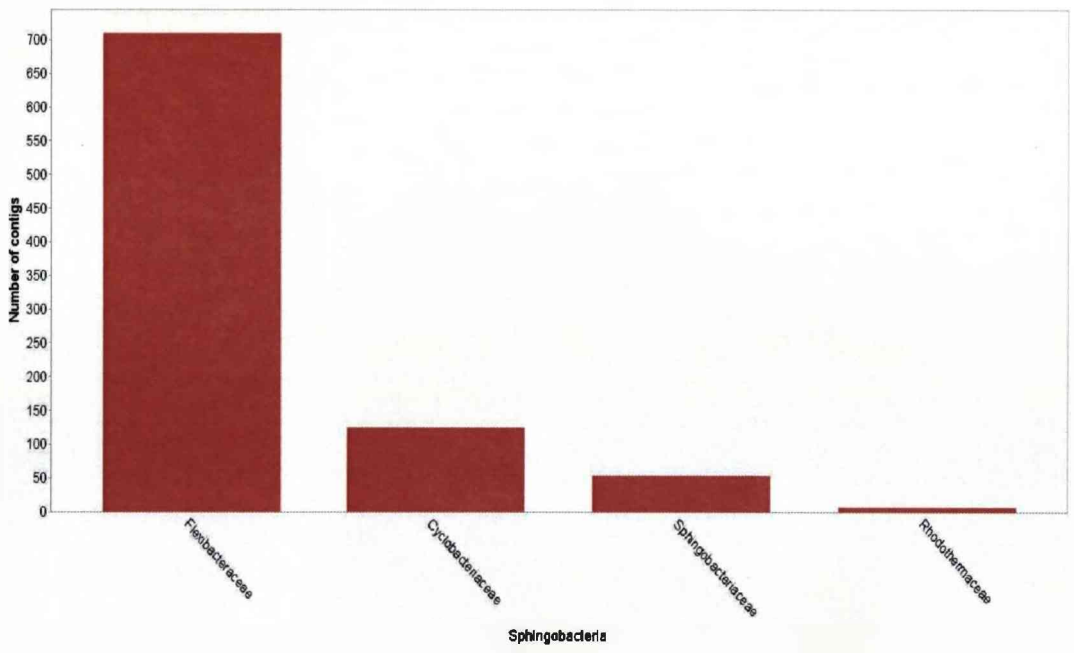


Figure 3.6Taxonomic diversity of contigs at the Phylum of *Bacteroidetes*

Contigs (indicated as number of reads) distribution within the classes of (A) *Flavobacteria* and (B) *Sphingobacteria* as computed by MEGAN.

Resolution of the taxonomic tree down to the species level revealed significant diversity within the *Gammaproteobacteria*, with 41 species represented (Figure 3.7). Several species were abundant with at least forty contigs assigned to each. Notably these all have a full reference genome sequence in GenBank (The numbers in brackets represent the number of assignments per taxon-followed by their GenBank accession number).

Cellvibrio japonicus (198) (GenBank: CP000934); *Pseudoalteromonas tunicata* (41) (AAOH00000000); *Pseudoalteromonas atlantica* (668) (CP000388); *Alteromonas macleodii* (73) (NC 011138); *Saccharophagus degradans* (742) (CP000282); *Colwellia psychrerythraea* (148)(CP000083) and *Hahella chejuensis* (79)(CP000155).

Cellvibrio japonicus is a saprophytic soil bacterium possessing representatives of 13 CBM families and 130 GHs sharing half of the plant cell wall degradative enzymes with that of *S. degradans* (Deboy *et al.*, 2008). *S. degradans*, a known marine polysaccharide-degrader capable of degrading several complex polysaccharides has had its cellulolytic system fully characterised (Taylor *et al.*, 2006), while, *A. macleodii, C. psychrerythraea, H. chejuensis*, all contain predicted cellulase genes on their genomes (http://www.uniprot.org). Members of the *Pseudoalteromonas* genus are regularly studied for their biofilm forming and antifouling properties (Egan *et al.*, 2002; Thomas *et al.*, 2007) and it is thought that members of the antifouling subgroup of the *Pseudoalteromonas* genus are capable of colonising and utilising cell wall cellulose from higher marine organisms (Skovhus *et al.*, 2007). Production of a number of inhibitory compounds by *P. tunicata* has been identified providing evidence for competitiveness in high density communities (Thomas *et al.*, 2008). It has also been observed that cellulose can induce the expression of the MSHA (type IV) pili in *P. tunicata* increasing attachment (Skovhus *et al.*, 2007) and therefore promoting the ability to bind to insoluble cellulosic substrates.

The *Alphaproteobacteria* in the dataset also appeared to be diverse, with 33 species represented with five or more contigs to each node (Figure 3.8). Apart from one of six *Roseobacter* species, which has forty-six assigned contigs there is no

predominant species represented in the dataset assigned taxonomically to the *Alphaproteobacteria.*

Twenty-nine species could be classified at the species level within the *Bacteroidetes,* with twelve species assigned forty contigs or more (Figure 3.9). Although this is not as diverse as the *Gamma* and *Alpha-proteobacteria,* there are still dominant species present identified with bacterial strains whose genomes have been sequenced: *Polaribacter sp.* (47) AANA00000000); *Flavobacterium johnsoniae* (80) (CP000685); *Psychroflexus torques* (62)(AAPR00000000); *Leewenhoekiella blandensis* (50)(AANC00000000); *Dokdoria donghaensis* (117) (AAMZ00000000); *Kordia algicid*a (107) (ABIB00000000); *Cytophaga hutchinsonii* (425) (CP00385); *Microscilla marina* (205) (AAWS00000000); *Algoriphagus sp* (125) (AAXU00000000); *Pedobacter sp* (54)(ABCM00000000). There were also three unclassified *Flavobacteriales* species collectively represented by 306 contigs. Members of the *Bacteroidetes* have been shown to exhibit gliding motility, including *F. Johnsoniae* and *C.hutchinsonii* (Liu *et al.,* 2007). Potentially this would enable cells to colonise other areas of the substrate. *Polaribacter sp* Med 152 is associated with diatom blooms and contains a large number of genes involved in surface or particle attachment and polymer degradation (Gonzalez *et al.,* 2008). *Psychroflexus torques* is an algal epiphyte (Bowman *et al.,* 1998) isolated from Antarctica sea ice, *Pedobacter* spp. posess carbohydrate scavenging activities and *Leewenhoekiella blandensis* is thought to be important in the cycling of nutrients (http://www.ncbi.nlm.nih.gov/nuccore/AANC00000000). Therefore, there is an abundance of bacteria within the *Bacteroidetes* suggested to be present here and they are emerging as influential community members in the degradation of organic materials (Bauer *et al.,* 2006). The biofilm community of the cellulose bait appears to contain an abundance of organisms with gliding motility that would provide the opportunity to find new areas of colonisation on polymeric complexes.

One of the problems inherent in the characterisation of this dataset based on reference databases, is that assignment of a number of contigs to a particular sequenced species detracts from the fact that many strains of a species, or indeed

many species within a genus, are present but not represented in the reference database. This is evident with the number of *Roseobacter* species assigned contigs (6) (Figure 3.8) where six of the eight *Roseobacter* species with complete genome sequences are undergoing assembly (March 2009) (www.ncbi.nlm.nih.gov/sites/entrez?db=genome). If only one species had a reference genome possibly all assignments would be placed at that taxon reducing the diversity present in the biofilm community. This is an important factor in annotation using limited reference databases.

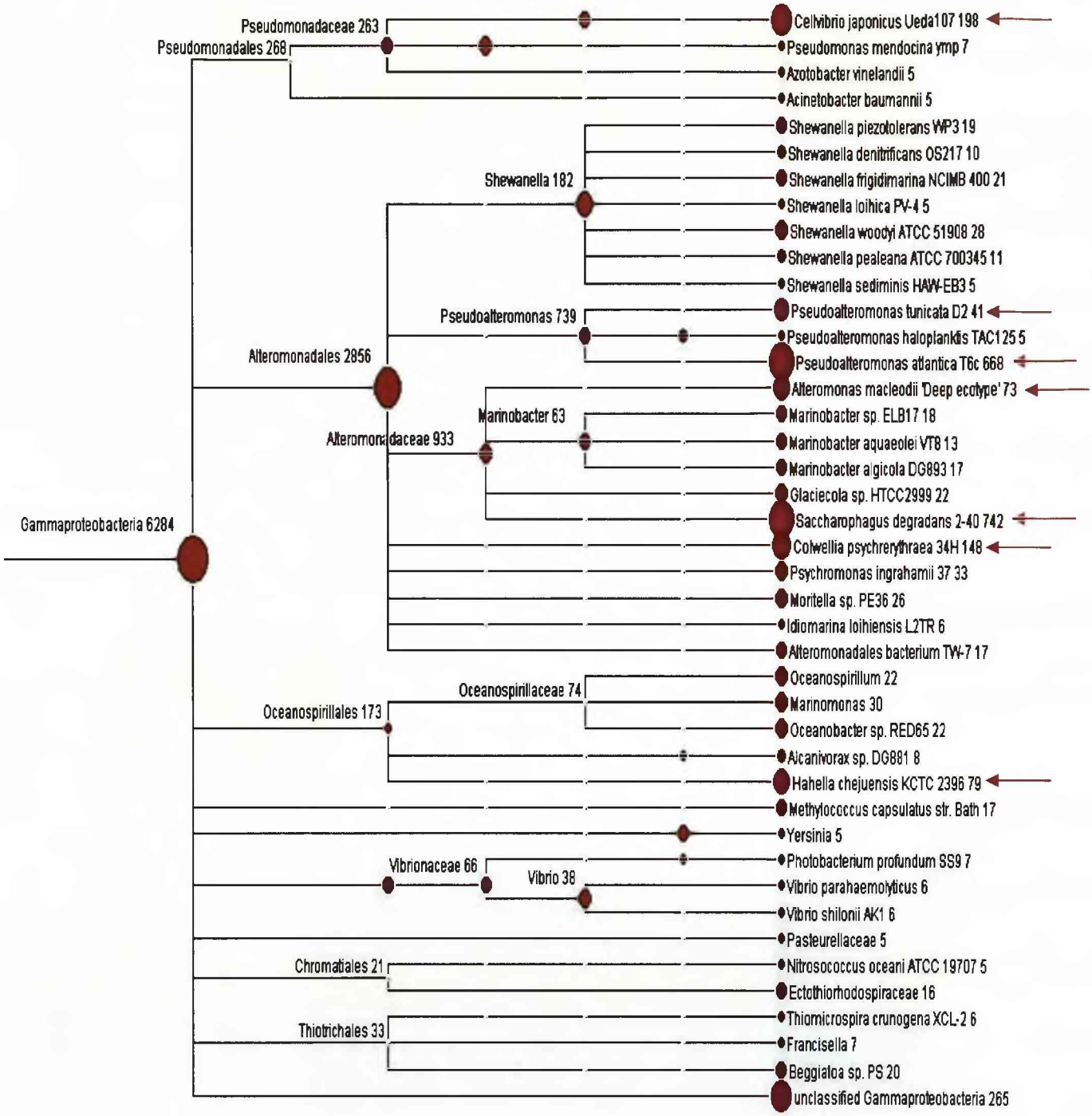


Figure 3.7 Taxonomic distribution of the bacterial species belonging to the *Gammaproteobacteria* found in the Irish Sea cellulose biofilm DNA 454 assembled contigs

Analysis of 26,860 assembled contigs based on blastx comparison against the NCBI nr database and taxonomically classified by MEGAN. The numbers of contigs assigned to each taxon is labelled (last number on each node). The seven arrows indicate the species most representing the *Gammaproteobacteria* in the dataset.

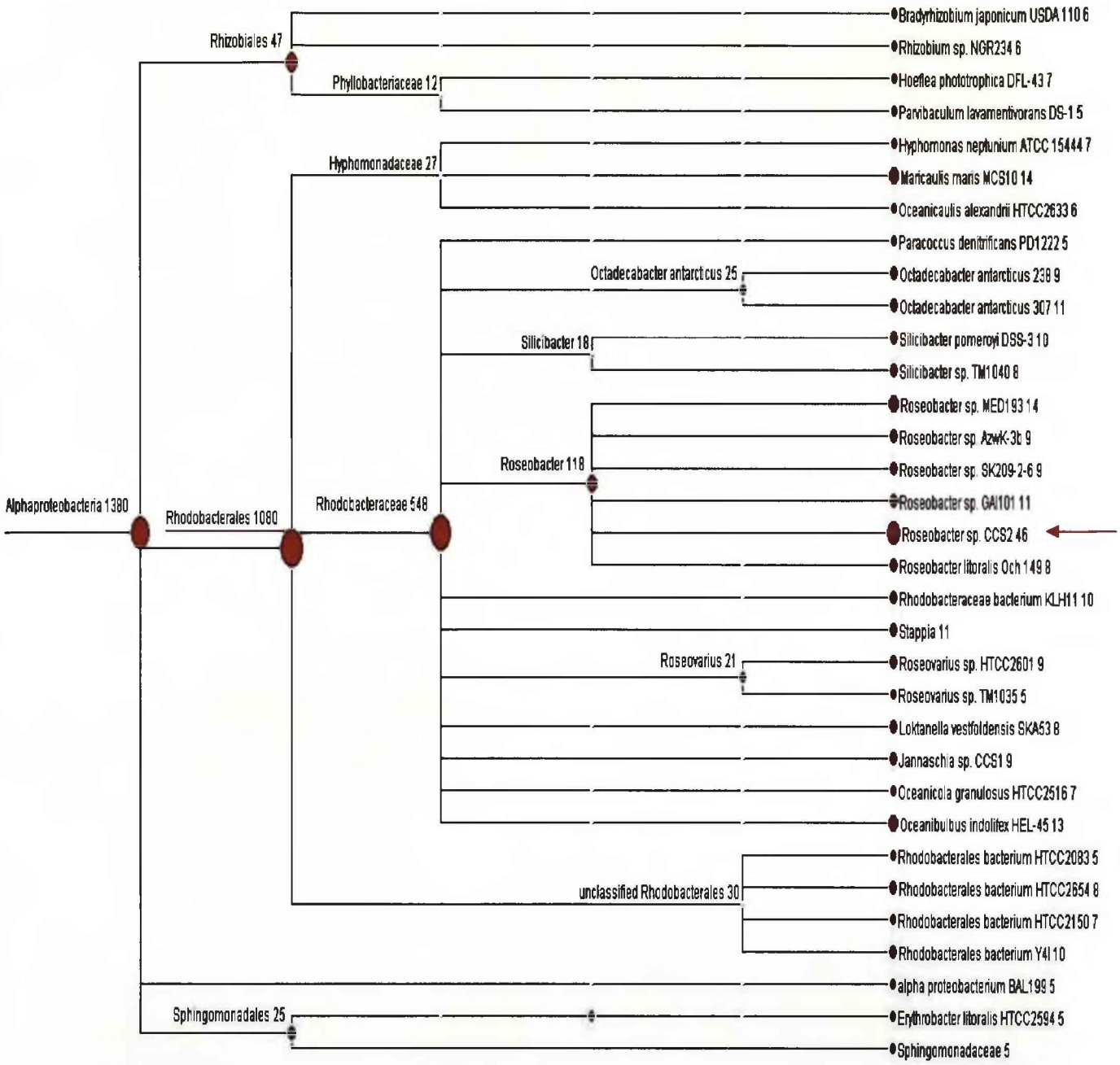


Figure 3.8 Taxonomic distribution of the bacterial species belonging to the *Alphaproteobacteria* found in the Irish Sea cellulose biofilm DNA 454 assembled contigs

Analysis of 26,860 assembled contigs based on blastx comparison against the NCBI nr database and taxonomically classified by MEGAN. The numbers of contigs assigned to each taxon is labelled (number at the end of each node). The arrow shows the most abundantly represented species which is one of six *Roseobacter* spp. revealed by 46 contigs.

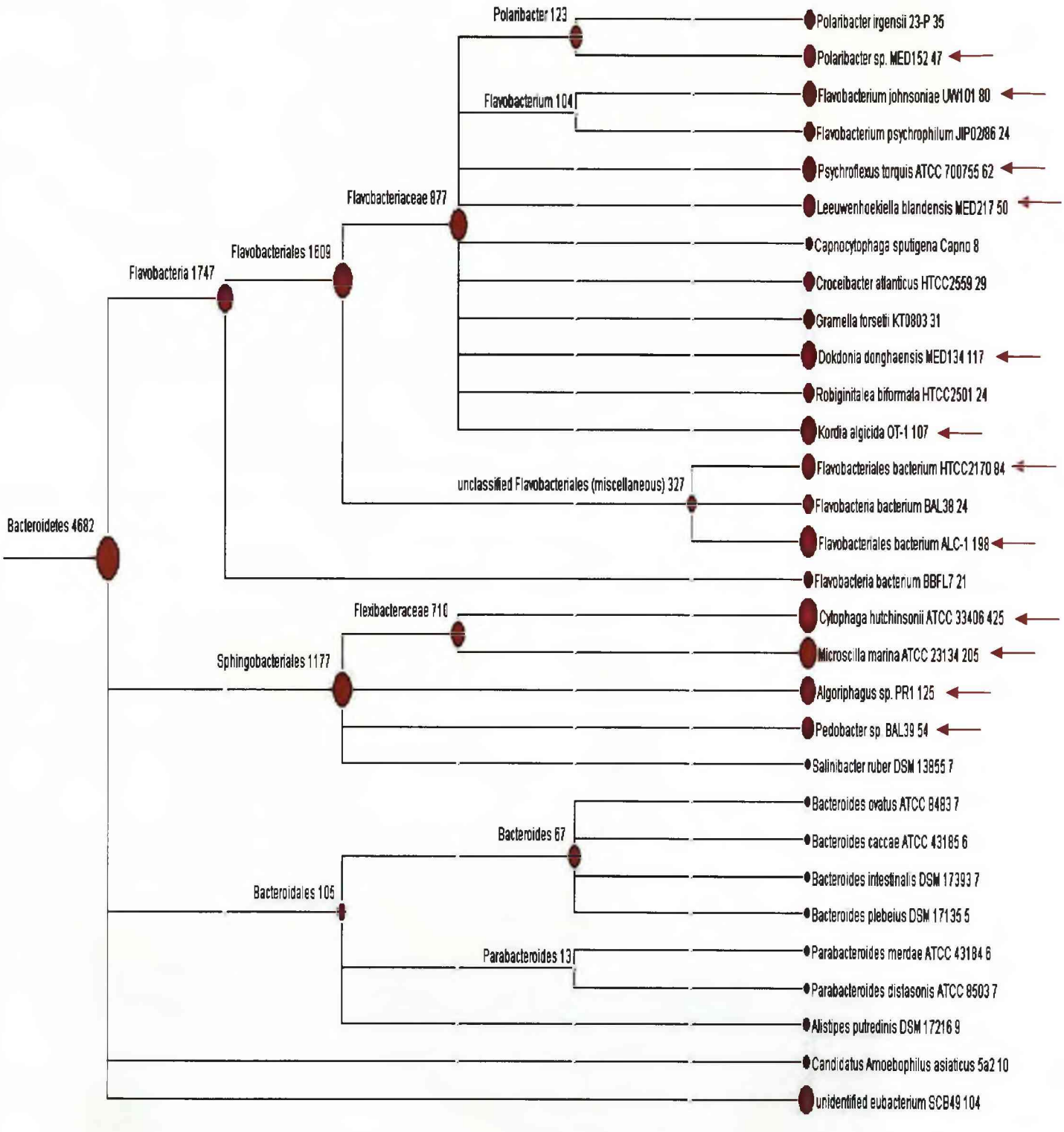


Figure 3.9 Taxonomic distribution of the bacterial species belonging to the *Bacteroidetes* found in the Irish Sea cellulose biofilm DNA 454 assembled contigs

Analysis of 26,860 assembled contigs based on blastx comparison against the NCBI nr database and taxonomically classified by MEGAN. The numbers of contigs assigned to each taxon is labelled (last number of each node). The arrows show the most abundant species represented of the *Bacteroidetes*.

Analysis of the closest taxonomic matches of the sequences is a feature of the MEGAN program. This enables some estimation of key traits of the microbial community. Figure 3.10 represents the number of classified species having an indicated property, and is based on the NCBI "prokaryotic attributes table", that describes cellular features, environmental and, temperature relationships, and pathogenicity. Notably the species assignments here comprise large numbers of aerobes (3307), organisms associated with the aquatic environment (3218), mesophilic organisms (3448), Gram negative bacteria (3921) and motile taxa (3474). This overall classification is in keeping with a metagenome dataset derived largely from a marine heterotrophic bacterial community.

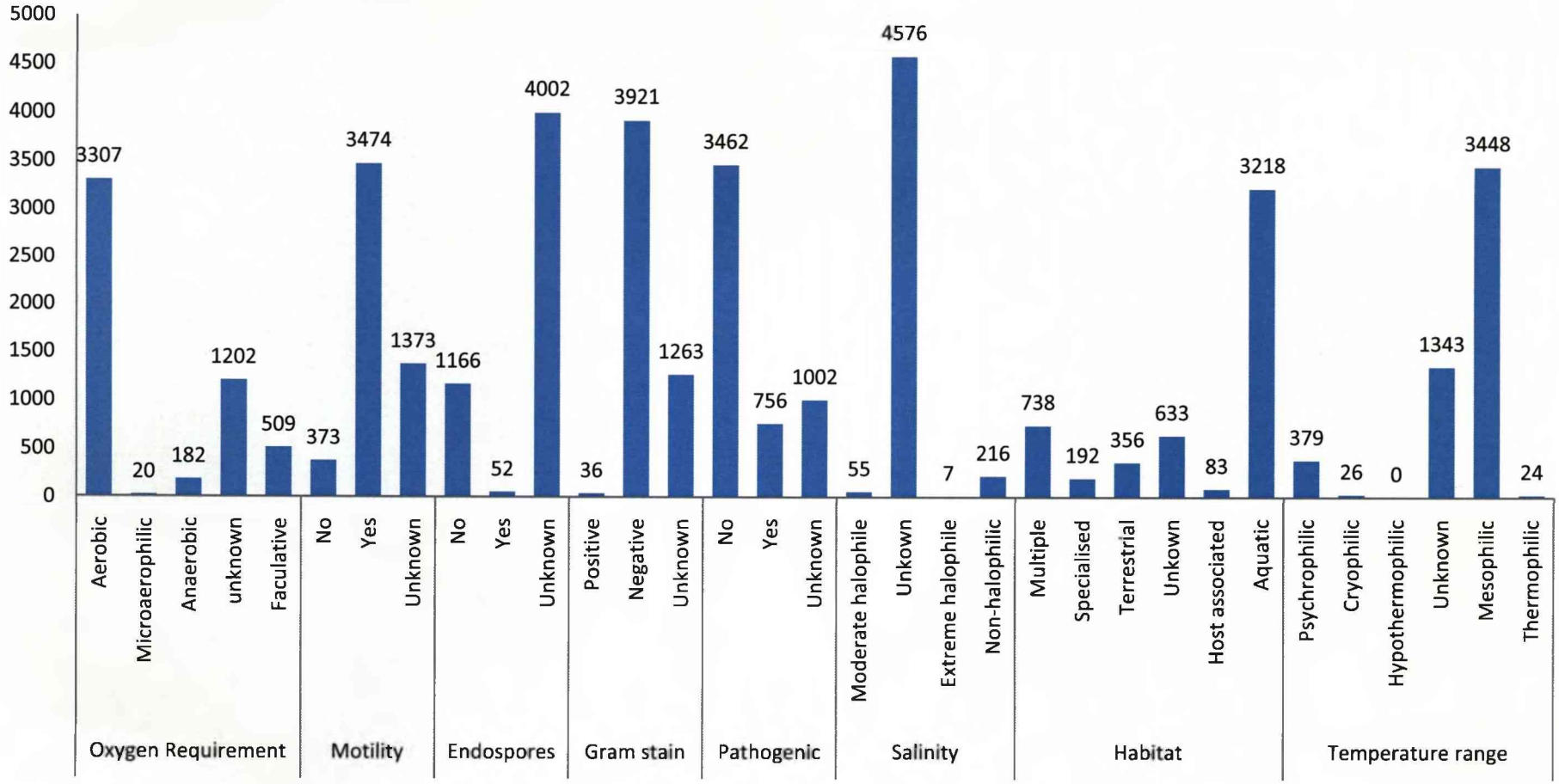


Figure 3.10 Microbial attributes associated with contigs of the Irish Sea cellulose biofilm DNA 454 assembled dataset as calculated by MEGAN.

Attributes are based on the NCBI "prokaryotic attributes table" as calculated by MEGAN. The number of contigs assigned to each attribute category is represented by the number of entries

3.3.3 MG RAST analysis of the Irish Sea cellulose biofilm DNA 454 assembled contigs

The Liverpool Bay assembled dataset was subjected to automated annotation using the MG-RAST (Meta Genome Rapid Annotation using subsystem technology) server at the Argonne National Library (http://metagenomics.nmpdr.org), using subsystem-based annotation based on the SEED database (Meyer *et al.*, 2008).

Figure 3.11, highlights overall assessment of the dataset provided by MG-RAST analysis, including the distribution of contigs in relation to GC content and the size distribution of the Irish Sea 454 pyrosequenced assembled contigs, notably the majority are under 1 kb in length.

All assembled contigs were compared to the SEED-nr database using blastx, where potential protein encoding genes (PEGs) are used for taxonomic affiliation. Figure 3.12 shows the protein based taxonomic classification of those PEGs matched with that of the SEED database with an E-value less than 0.01. This e-value was chosen from the selection on the MG-RAST website to retrieve potentially coding elements (Meyer *et al.*, 2008) and provides an overall view of a microbial community based on annotation solely in the reference database. Of the 26,860 contigs in the dataset, 16,742 (62%) could be classified with the overwhelming majority being assigned to the *Bacteria* (16,490). A small number of contigs represented the *Eukarya* (142), Viruses (40) and the *Archaea* (70).

As with the MEGAN analysis, the dataset is dominated by the *Proteobacteria* (63%) and the *Bacteroidetes/Chlorobi* group (31 %) (Figure 3.12).

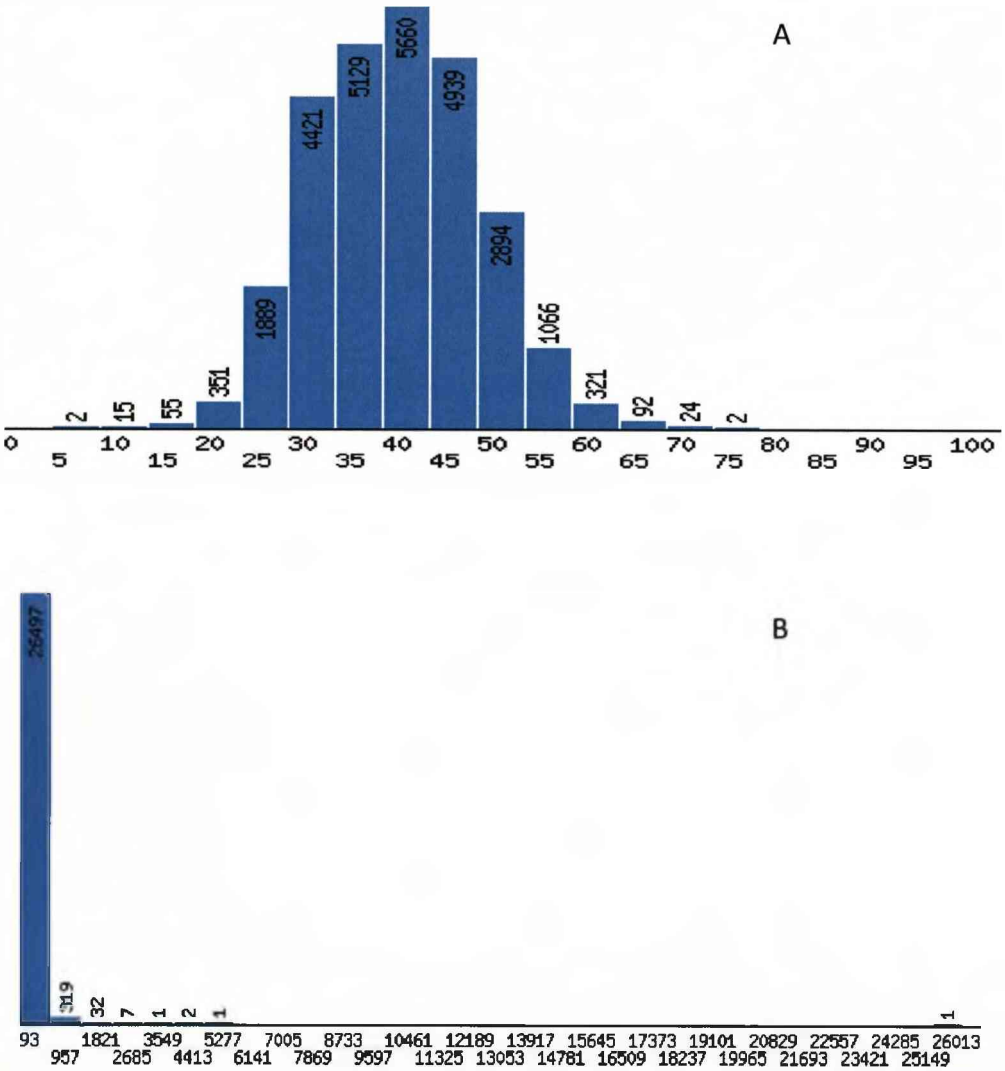


Figure 3.11 MG-RAST based Overview of the Metagenome sequences

A, displays the distribution of the GC percentage for the metagenome sequences. Each bar represents the number of sequences in that GC percentage range. **B**, shows the distribution of sequence lengths for this metagenome. Each bar represents the number of sequences for a certain length range.

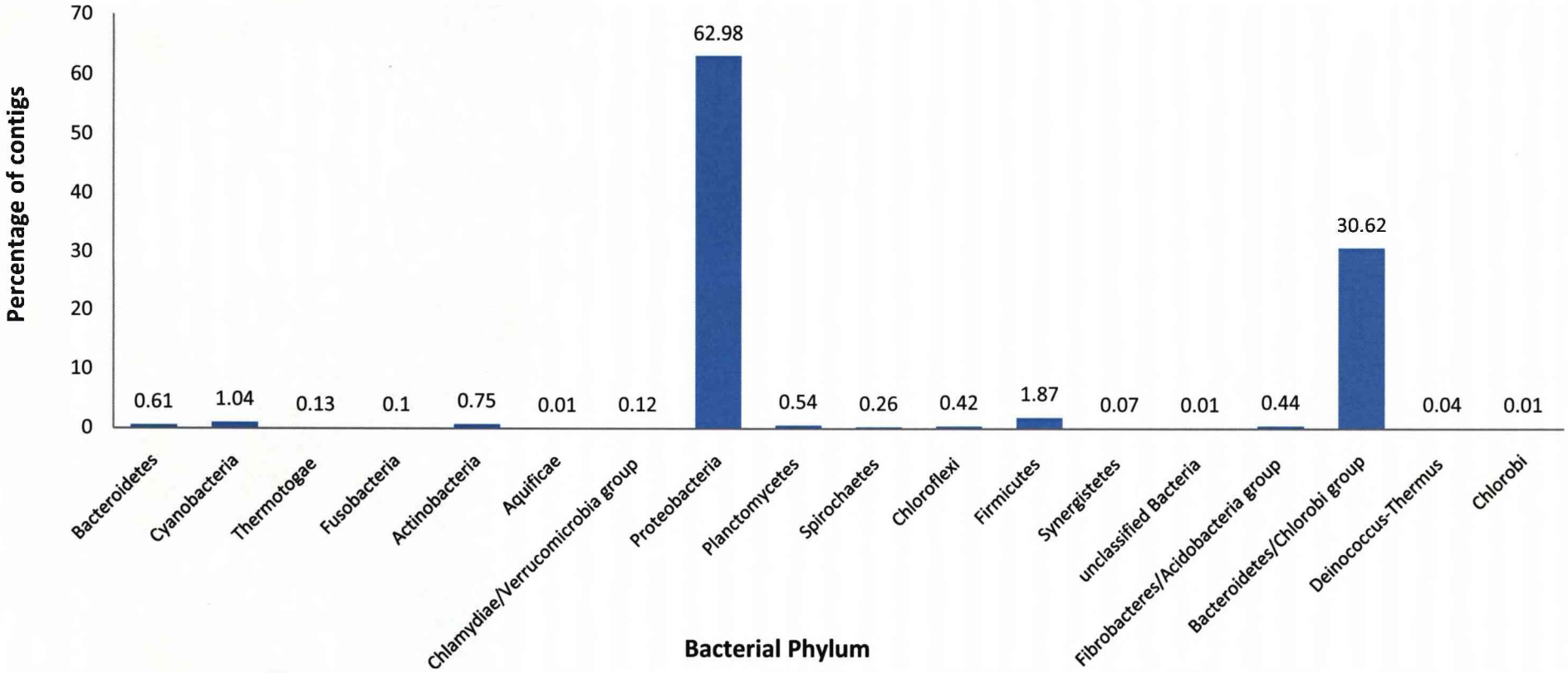


Figure3.12 Taxonomic Distribution of contigs in the Irish Sea cellulose biofilm DNA 454 dataset as computed by MG-RAST.

Percentages are of those contigs affiliated with phyla within the domain *Bacteria* following a blastx comparison against the SEED database computed by MG-RAST.

The dataset was also annotated by similarity matching against SEED subsystems. Comparison of the Irish Sea dataset to SEED subsystems highlights homology of sequences in the dataset to genes in known subsystems (categories based on annotation of reference genomes) to provide metabolic reconstruction. Of the 26,860 contigs in the dataset, 10,620 (39.5%) could be assigned a functional role within a metabolic pathway by matching to proteins in the SEED subsystems, using an e-value cut off of 0.01. Figure 3.13 shows details of the classification to divisions within the subsystems. Notably, clustering based subsystems (including a wide array of functions such as cell division) represent the largest proportion (15.42%). The carbohydrate subsystem (carbohydrate metabolism and uptake genes) is also highly represented (11.39%).

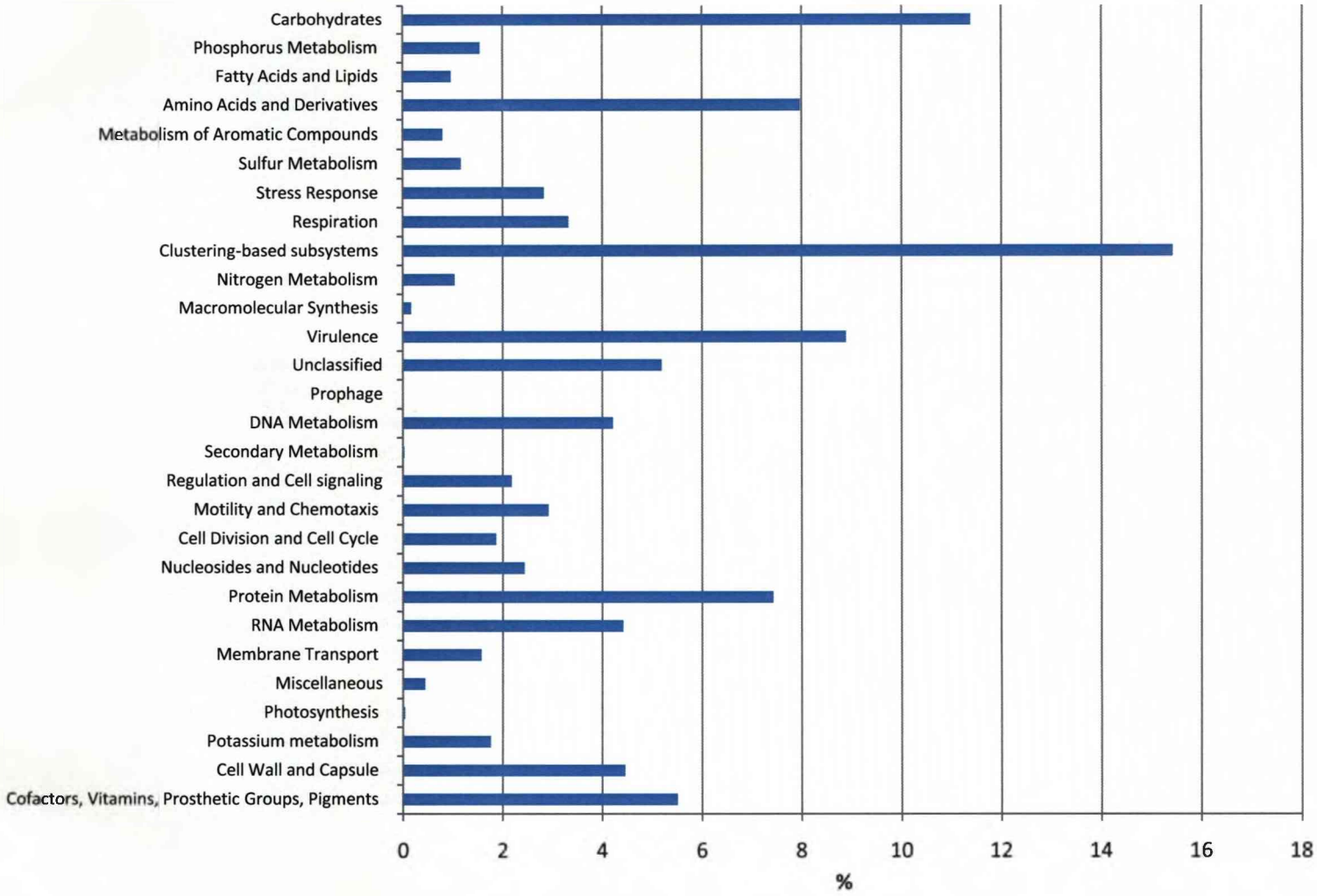


Figure 3.13 SEED subsystem composition of contigs assembled from the Irish Sea cellulose biofilm DNA 454 dataset. Percentages are expressed as the number of contigs assigned to a subsystem category, divided by the total number of contigs assigned to subsystems as calculated by MG-RAST.

3.3.4 Phylogenetic analysis of the Irish Sea 454 dataset

Phylogenetic analysis based on highly conserved genes is routinely used in molecular microbial ecology; for bacteria, usually the 16S rRNA gene is amplified from environmental DNA using universal primers, cloned and sequenced to catalogue the distribution of species in an environment (for review see Kirchman, 2008). With MG-RAST all contigs in the dataset are also compared by blastn to a number of rRNA sequence databases.

Table 3.2 Shows hits for the dataset against the Greengenes database using an e-value cut off of 0.01 and a recommended minimum alignment length of 50 bp (Meyer *et al.*, 2008). Eighteen contigs revealed hits against the database, with fourteen affiliated with Bacteria - of which 4 are affiliated to the *Bacteroidetes* (*Flavobacterium; Ulvibacter; Roseivirga* and unclassified *Flexibacteraceae*) and 10 to the Proteobacteria (*Sulfitibacter;* Unclassified_*Rhodobacteraceae* (3); *Glaciecola* (3); *Teredinibacter; Cellvibrio* and Unclassified_*Gammaproteobacteria*).

Alignment length of the contigs range from 85 bp for contig 16635, which has 100 % identity with a *Sulfitobacter* sp. to contig 00061 which has an alignment length of 892 bp and a 90% identity to a *Roseivirga* sp. (table 3.2). Three contigs show matches with the 16S rRNA gene of members of the genus *Glaciecola*, contig 25574, 26574 and 26860 with an alignment length of 235 bp, 526 bp and 151 bp respectively at 99-100% identity (table3.2). There are no matches to genera belonging to phyla outside the *Proteobacteria* and *Bacteroidetes* phyla in the Greengenes database and using the parameters described above.

Table 3.2 Comparison of the Irish Sea cellulose biofilm DNA 454 dataset to the Greengenes 16S rRNA gene database

Sequence ID	Alignment Length	E-value	% Identity	Bit Score	Fragment (Start - End)	Taxonomy Assignment	Best Hit ID	
contig00070	247	1.26E-81	93	305	738 - 983	*Cellvibrio*	98921	Proteobacteria
contig12909	577	0.00E+00	93	858	209 - 784	unclassified_*Gammaproteobacteria*	151615	
contig16635	85	1.00E-40	100	168	1-85	*Sulfitobacter*	165917	
contig25574	235	2.51E-130	100	466	50 - 284	*Glaciecola*	108683	
contig26360	150	2.51E-51	93	202	1 - 149	unclassified_*Rhodobacteraceae*	142124	
contig26572	162	7.94E-70	96	264	4 - 163	unclassified_*Rhodobacteraceae*	70710	
contig26573	526	0.00E+00	99	1003	1 - 526	*Glaciecola*	80428	
contig26707	130	2.51E-63	99	242	1 - 130	unclassified_*Rhodobacteraceae*	113926	
contig26820	696	0.00E+00	93	1015	6 - 701	*Teredinibacter*	144812	
contig26860	151	1.26E-80	100	299	1 - 151	*Glaciecola*	170839	
contig00061	892	0.00E+00	90	1094	1988 - 2879	*Roseivirga*	102384	Bacteroidetes
contig26228	139	2.51E-66	98	252	1 - 138	*Flavobacterium*	154970	
contig26430	241	5.01E-130	99	464	1 - 241	*Ulvibacter*	80102	
contig26765	436	5.01E-140	89	500	60 - 495	unclassified_*Flexibacteraceae*	2577	

Best Hit ID refers to the ProkMSA reference ID.

Comparison of the dataset with the RDP database using an e-value cut off of 0.01 and a recommended minimum alignment length of 50 bp gave matches for 32 contigs. All are affiliated to *Bacteria*, 5 to the phylum *Bacteroidetes* and 21 to the *Proteobacteria*. 6 contigs cannot be affiliated higher than the level of Bacteria (Table 3.3). The shortest alignment length is 52 bp for that of contig 26605 with 94 % identity to an unclassified member of the *Oceanospirillales*. The largest alignment is 892 bp for contig 00061 with 90% identity to an unclassified member of the *Flexibacteraceae*. Five contigs assigned to the *Proteobacteria* can be affiliated at the genus level (*Sulfitobacter; Colwellia; Glaciecola* (2) and *Cellvibrio)* (Table 3.3) whilst only two affiliate to *Bacteroidetes* (*Cellulophaga* and *Psychroserpens*) (Table 3.3). Four contigs (14917, 26546, 04032 and 05558) share the top hit of S000444465_Unclassified *Gammaproteobacteria* and two contigs (26605 and 26650) share a top hit of S000141133_Unclassified *Oceanospirillales*. This could infer that the contig sequences originate from strains of the same species. It is also probable that those contigs in the data set that originate from 16S rRNA genes are from species that are predominant in the community.

Although specific genera are recognised to be present in the dataset by analysis against the RDP and greengenes databases (i.e. *Glaciecola, Colwellia, Sulfitobacter, Cellvibrio, Cellulophaga and Psychroserpens*), the results comprise a high proportion of unclassified members of the Proteobacteria and *Bacteroidetes*. This is in accordance with the view that a plethora of bacteria have not only remained as yet uncultivated, but that their biochemical properties and function in the marine ecosystem have yet to be elucidated.

Using both databases, no matches were identified for the full 16S rRNA gene (~ 1500 bp) due to the size of the assembled contigs. This could result in incorrect assignments but nevertheless provides a sound basis for inferring the presence of key groups of bacteria in the marine cellulose colonising community. Additionally a significantly greater proportion of contigs were matched to the Proteobacteria using the RDP analysis. However observing the hits closely a large number of these contains alignment lengths of less than 100 bp.

Table 3.3 Comparison of the Irish Sea cellulose biofilm DNA 454 dataset to the Ribosomal Database Project (RDP) 16S rRNA gene database

Sequence ID	Alignment Length	E-value	% Identity	Bit Score	Fragment (Start - End)	Taxonomy Assignment	Best Hit ID	
contig12909	516	0.00E+00	95	829	267 - 781	unclassified_*Oceanospirillales*	S000105828	Proteobacteria
contig14917	63	7.94E-21	95	101	120 - 182	unclassified_*Gammaproteobacteria*	S000444465	
contig16635	97	7.94E-41	96	168	1 -97	*Sulfitobacter*	S000455479	
contig20563	77	1.00E-26	94	121	8 - 84	unclassified_*Gammaproteobacteria*	S000379224	
contig24279	241	1.00E-112	96	406	1 - 241	*Colwellia*	S000412159	
contig25574	244	6.31E-136	100	484	41 - 284	*Glaciecola*	S000363899	
contig25882	75	1.00E-27	96	125	310 - 384	unclassified_*Xanthomonadaceae*	S000361337	
contig26360	150	5.01E-49	92	194	1 - 149	unclassified_*Rhodobacteraceae*	S000346623	
contig26546	59	1.26E-06	86	54	94 - 152	unclassified_*Gammaproteobacteria*	S000444465	
contig26557	113	3.16E-21	87	105	16 - 128	unclassified_*Xanthomonadaceae*	S000016619	
contig26572	162	5.01E-70	96	264	4 - 163	unclassified_*Rhodobacteraceae*	S000321074	
contig26573	526	0.00E+00	99	1003	1 - 526	*Glaciecola*	S000394887	
contig26605	52	3.16E-12	94	71	48 - 99	unclassified_*Oceanospirillales*	S000141133	
contig26650	78	2.51E-17	89	91	2 - 79	unclassified_*Oceanospirillales*	S000141133	
contig26707	130	2.51E-63	99	242	1 - 130	unclassified_*Rhodobacteraceae*	S000493479	
contig26820	599	0.00E+00	96	1037	6 - 604	unclassified_*Gammaproteobacteria*	S000486766	
contig26860	151	1.00E-80	100	299	1 - 151	unclassified_*Gammaproteobacteria*	S000402009	
contig00070	247	1.26E-81	93	305	738 - 983	Cellvibrio	S000401934	
contig03976	77	2.51E-31	97	137	69 - 145	unclassified_*Enterobacteriaceae*	S000022545	
contig04032	57	2.51E-14	94	81	113 - 168	unclassified_*Gammaproteobacteria*	S000444465	
contig05558	59	1.26E-06	86	54	53 - 111	unclassified_*Gammaproteobacteria*	S000444465	
contig26228	139	2.51E-66	98	252	1 - 138	unclassified_*Flavobacteriaceae*	S000486614	Bacteroidetes
contig26430	241	5.01E-130	99	464	1 - 241	*Cellulophaga*	S000394819	
contig26563	204	1.00E-90	96	333	1 - 202	*Psychroserpens*	S000396985	
contig26765	436	5.01E-140	89	500	60 - 495	unclassified_*Sphingobacteriales*	S000414962	
contig00061	892	0.00E+00	90	1094	1988 - 2879	unclassified_*Flexibacteraceae*	S000425862	

Best Hit ID refers to the RDP reference

3.3.5 Glycosyl Hydrolase analysis of the Irish Sea 454 pyrosequencing dataset

FastA versions of glycosyl hydrolase (GH) domain families containing proteins identified as endoglucanases and chitinases were downloaded from the Pfam website (January, 2009) and a database constructed (Table 3.1). Pfam was chosen as it provides a collection of curated sequences belonging to protein families (www.sanger.ac.uk/Software/Pfam/). All 26,860 pyrosequencing derived contigs were used as queries in a blastx (all against all) search against the constructed glycosyl hydrolase database. The method was used to pre-process or 'hook out' those contigs with sequence homolgy to endoglucanase and chitinase domains in the Pfam dataset, reducing significantly the number of contigs to be further analysed. Contigs which produced hits as a result of the search were individually subjected to a blastx comparison against the NCBI-nr database. As a result of this a small number of contigs that initially had similarity to the GH database (constructed with protein sequences downloaded from Pfam) subsequently had top hits to proteins other than GH's when compared to a different reference database (NCBI-nr), such as contig 00952 which produced a top hit to an NADH-ubiquitone oxidoreductase of *Microscilla marina* (Table 3.4). In addition the families chosen contain proteins of other fuctions than endoglucanase and chitinase, making it possible to pick out proteins such as Mannanase. As comparisons are made between metagenome sequences and those of known GH families, there is of course the potential to miss new GH families, should any exist. However due to motif and fold similarities between different families (clans) this would only be a significant problem for any such families that had completely new folding patterns. The total number of query sequences which showed homology to GH sequences in the database according to the the parameters set (compromise between speed and sensitivity) was 116 (Table 3.4). Those hits that were most frequent were proteins from *Saccharophagus degradans, Cytophaga hutchinsonii* and *Cellvibrio japonicus*. *S. degradans* is probably the best studied marine cellulose degrading species to date (Taylor *et al.*, 2006) whilst the latter two are known cellulolytic bacteria isolated from soil (Xie *et al.*, 2007; Deboy *et al.*, 2008). The limited representation of marine cellulose degrading bacteria in reference databases may be the reason for the

frequency of hits because these species have representative sequenced genomes and well characterised cellulase systems. Multiple contigs also hit the same protein for example contigs 00162 & 01065 both have a top hit against an endoglucanse (YP 678840.1) from *C. Hutchinsonii.* Both contigs are quite large (421 bp & 773 bp respectively) with good e-values (4e-19 & 1e-57 respectively).

The large number of genes (116) identified as potentially involved in the cellulose degrading process, provide evidence of a microbial community functionally adapted to the colonisation and degradation of complex polysaccharides, and a feature of the use of *in situ* colonised cellulose as the source of the metagenome dataset. It should also be remembered that presented is the hits to a limited number of the total GH reference database and cellulases with exo-acting mechanisms and β-glucosidases may have been missed. Additionally comparison to the Carbohydrate binding module (CBM) families and to cellulosomal modules such as cohesions and dockerins may provide further insight into the genetic repertoire involved in cellulose degradation in this environment.

Table 3.4 Results of BLASTX comparison of Irish Sea dataset against a customised Glycosyl Hydrolase database.

Contig	Length (bp)	Top hit	Organism hit	e-value	% Identity	% similarity
00014	343	Hypothetical protein (NP 285551.1)	*Deinococcus radiodurans*	3e-9	50 (16/32)	68 (22/32)
00162	421	Endoglucanase (YP 678840.1)	*Cytophaga hutchinsonii*	4e-19	56 (42/74)	74 (55/74)
00226	460	Glycosyl Hydrolase family 8 (YP 680017.1)	*Cytophaga hutchinsonii*	4e-45	67 (102/152)	76 (116/152)
00230	176	Mannanase (BAA25878.1)	*Baciilus circulans*	3e-12	57 (32/56)	76 (42/56)
00357	101	B-1,4-Cellobioside (YP 001829271.1)	*Xylella fastidiosa*	0.025	55 (15/27)	81 (22/27)
00385	190	Hypothetical protein (YP 001915259.1)	*Xanthomonas oryzae*	8e-22	72 (45/62)	87 (54/62)
00952	222	NADH-ubiquitone oxidoreductase (ZP01690931.1)	*Microscilla marina*	4e-36	91 (68/74)	97 (72/74)
01065	773	Endoglucanase (YP 678840.1)	*Cytophaga hutchinsonii*	1e-57	45 (121/267)	63 (168/267)
01404	186	Cold shock DNA binding protein	*Saccharophagus degradans*	1e-25	91 (54/59)	91 (54/59)
01678	312	2-isopropylmalmate synthase (YP 527962.1)	*Saccharophagus degradans*	4e-28	53 (55/102)	67 (69/102)
01825	418	Hypothetical Protein (ZP01255148.1)	*Psychroflexus torques*	9e-11	30 (44/142)	49 (69/142)
01895	167	Exo-oligoxylanase (BAF49077.1)	*Paenibacillus sp*	3e-08	55 (26/47)	72 (34/47)
02283	216	Endoglucanase I (ABU45498.1)	*Fibrobacter succinogenes*	4e-08	39 (20/51)	62 (32/51)
02598	373	Cellulase (BAB79288.1)	*Pseudomonas sp*	1e-30	46 (58/124)	66 (83/124)

02656	302	Endo 1,4 β-D glucanase (ABB51610.1)	Uncultured bacteria	3e-23	53 (52/98)	67 (66/98)
02702	217	Glycosyl Hydrolase family 16 (YP563606.1)	Shewanella denitrificans	6e-25	73 (52/71)	78 (56/71)
02758	590	Cellulase (YP528708.1)	Saccharophagus degradans	2e-19	47 (34/72)	58 (42/72)
02917	389	Endo 1,4 β-glucanase (ABB51609.1)	Uncultured bacterium	2e-43	68 (81/119)	82 (98/119)
02925	313	Endoglucanase (YP 678336.1)	*Cytophaga hutchinsonii*	5e-48	80 (84/104)	87 (91/104)
03316	196	Cellulase (YP 526110.1)	*Saccharophagus degradans*	3e-15	55 (35/64)	71 (46/64)
03941	831	PDK repeat containing protein	*Chitinophaga pinensis*	0.014	29 (40/134)	45 (61/134)
04098	301	Endo-1,4-β D glucanase (ABB51609.1)	Uncultured bacterium	2e-22	52 (54/102)	66 (68/102)
04105	165	B 1,4-cellobiosidase (YP 002203140.1)	*Streptomyces sviceus*	4e-05	45 (28/62)	54 (34/62)
04179	608	B-glucanase precursor (ZP01719353.1)	*Algoriphagus sp*	1e-49	55 (85/153)	75 (116/153)
04307	235	Hypothetical protein (ZP 02072724.1)	*Bacteroides uniformis*	6e-7	73 (19/26)	84 (22/26)
04413	751	Dihydrolipoamide dehydrogenase (ZP 01691494.1)	*Microscilla marina*	2e-50	73 (99/134)	81 (109/134)
04615	190	Bi-functional xylanase/esterase (YP677852.1)	*Cytophaga hutchinsonii*	8e-17	65 (41/63)	77 (49/63)
04687	268	Endo 1,3-1,4 β-glucanase (YP 001615240.1)	*Sporangium cellulosum*	1e-20	51 (45/88)	64 (57/88)
04690	450	GH 8 B-glycosidase, CBM9	*Cytophaga hutchinsonii*	7e-55	68 (102/148)	81 (121/148)
04786	176	Cellulase (ZP 01107011.1)	*Flavobacteriales bacterium*	2e-08	52 (30/57)	66 (38/57)
04801	125	Thiamine monophosphate kinase (YP 528887.1)	*Saccharophagus degradans*	2e-05	65 (21/32)	78 (25/32)

05113	455	Putative cellulose Cel5D (YP 001983464.1)	*Cellvibrio japonicus*	9e-34	54 (65/120)	69 (83/120)
05185	375	Endo 1,4 β-glucanase precursor (AF208495.1)	*Pectobacterium chrysanthemi*	6e-36	56 (70/123)	77 (95/123)
05199	211	Cellulase (BAB79288.1)	*Pseudomonas sp* ND 137	5e-12	61 (43/70)	74 (52/70)
05226	241	Penicillin-binding protein (YP 268453.1)	*Colwellia psychrerythraea*	6e-28	74 (59/79)	87 (69/79)
05485	502	Endo 1,4 β-glucanase (AB51611.1)	Uncultured Bacterium	1e-44	60 (100/166)	69 (115/166)
05603	896	Regulatory Protein (YP526034.1)	Saccharophagus degradans	8e-86	54 (96/177)	68 (122/177)
05872	132	Endoglucanase (YP 678265.1)	Cytophaga hutchonsonii	2e-11	74 (32/43)	83 (36/43)
05967	225	Cellulase (CAF22221.1)	Uncultured bacterium	2e-12	70 (50/71)	76 (54/71)
06033	573	Endoglucanase-like protein	*Cytophaga hutchinsonii*	1e-24	37 (64/170)	51 (88/170)
06353	376	Predicted protein (XP 001215690.1)	*Aspergillus terreus*	5.3	36 (18/49)	55 (27/49)
06502	352	Putative cellulose Cel5D (YP 001983464.1)	*Cellvibrio japonicus*	4e-26	47 (52/109)	69 (76/109)
06707	184	Endoglucanase (YP678265.1)	*Cytophaga hutchinsonii*	2e-19	70 (42/60)	91 (55/60)
06721	688	GH 8 B-glycosidase, CBM 9 (YP 680303.1)	*Cytophaga hutchinsonii*	4e-98	70 (166/234)	82 (193/234)
06725	518	Putative chaperone (YP340674.2)	*Pseudoalteromonas haloplantktis*	4e-43	51 (51/100)	74 (74/100)
06951	185	Aminotransferase (ZP 02145308.1)	*Phaeobacter gallaeciensis*	2e-27	95 (58/61)	98 (60/61)
07103	391	2-isopropylmalate synthase (YP 527962.1)	*Saccharophagus degradans*	1e-16	39 (53/135)	52 (71/135)
07114	163	Cellulase (CAD61242.1)	Uncultured bacterium	5e-04	51 (15/29)	65 (19/29)

07415	604	B-glucosidase + CBM (YP 680303.1)	*Cytophaga hutchinsonii*	8e-57	71 (76/107)	81 (87/107)
07887	207	3-hydroxyacyl-CoA dehydrogenase (YP 527046.1)	*Saccharophagus degradans*	3e-09	72 (26/36)	88 (32/36)
08442	144	Endoglucanase B elgB (XP 001391969.1)	*Aspergillus niger*	0.37	63 (14/22)	86 (19/22)
08470	1001	Endoglucanase like protein (YP677947.1)	*Cytophaga hutchinsonii*	1e-31	54 (68/125)	66 (83/125)
08504	436	2-isopropylamate (YP527962.1)	*Saccharophagu degradans*	5e-04	41 (19/46)	58 (27/46)
08844	241	β-glucanase Precursor (ZP 01121077.1)	*Robiginitala biformata*	2e-26	69 (53/76)	85 (65/76)
08874	751	Succinyl-CoA synthetase (ZP 01121213)	*Robiginitalea biformata*	1e-55	75 (91/121)	85 (103/121)
08899	383	Endoglucanase (ZP 04358756.1)	*Chitinophaga pinensis*	8e-07	47 (29/61)	67 (41/61)
10055	241	Putative secreted β-Mannosidase (YP525540.1)	*Saccharophagus degradans*	5e-20	56 (45/80)	72 (58/80)
10093	241	Endo 1,4 β-glucanase (YP 001982113.1)	*Cellvibrio japonicus*	4e-15	52 (39/74)	70 (52/74)
10324	175	Glycosyl Hydrolase family 19 (YP 112524.1)	*Flavobacteria Phage*	3e-07	65 (23/35)	77 (27/35)
10599	241	Glycosyl Hydrolase family 16 (ZP 01612699.2)	*Alteromonadales* bacterium TW-7	2e-15	55 (43/78)	66 (52/78)
10691	661	Endoglucanase GH9 (YP 678265.1)	*Cytophaga hutchinsonii*	2e-12	35 (46/130)	56 (74/130)
11004	225	Endoglucanase (YP677892.1)	*Cytophaga hutchinsonii*	3e-21	64 (47/73)	78 (57/73)
11285	198	Serine protease (YP 271274.1)	*Colwellia psychrerythraea* 34 H	5e-8	50 (31/61)	67 (41/61)
11483	241	B 1,4-Cellobioside	*Xylella fastidiosa*	2e-08	58 (24/41)	75 (31/41)
11489	113	Acidic mammalian chitinase (XP 001868125.1)	*Culex quinquefasciatus*	3.1	56 (13/23)	82 (19/23)

12218	115	Endoglucanase (YP 677892.1)	*Cytophaga hutchinsonii*	1.4	50 (19/38)	63 (24/38)
12233	238	Endo 1,4 β-glucanase Cel 9B (YP 001982113.1)	*Cellvibrio japonicus*	9e-08	39 (30/76)	61 (47/76)
12453	166	Endo 1,3 β-glucosidase (YP 528590.1)	*Saccharophagus degradans*	4e-17	69 (38/55)	80 (44/55)
12578	557	Hypothetical protein (YP 002718093.1)	*Verrucomicrobiae bacterium*	2e-46	50 (45/90)	66 (60/90)
12863	241	Glycosyl Hydrolase family 5 (YP 002885352.1)	*Exiguobacterium sp*	6e-06	45 (27/60)	58 (35/60)
13597	459	Cellulase (YP 526110.1)	*Saccharophagus degradans*	1e-24	43 (43/100)	62 (62/100)
13913	342	Endo 1,4 β-glucanase Cel9B (YP 001982113.1)	*Cellvibrio japonicus*	2e-20	45 (51/113)	63 (72/113)
13936	135	Cellulase CelA (CAN91626.1)	*Sorangium cellulosum*	2e-07	59 (26/44)	75 (33/44
14052	331	CelA (CAN91626.1)	*Sorangium cellulosum*	1e-16	54 (32/59)	69 (41/59)
14077	238	Endo β1,4-glucanase Cel9B (YP 001982113.1)	*Cellvibrio japonicus*	3e-25	69 (55/77)	77 (61/77)
14328	241	Cellobiohydrolase II (BAG48183.1)	*Irpex lacteus*	3e-06	48 (25/52)	61 (32/52)
14405	271	DEHAZF16632p (CAG89465.2)	*Debaryomyces hansenii*	2e-10	32 (29/88)	62 (55/88)
14791	224	1,4 β-Mannanase (BAB79290.2)	*Pseudoalteromonas sp*	2e-06	45 (26/57)	64 (37/57)
15304	168	Putative Glycosyl Hydrolase (YP 001615952.1)	*Sorangium cellulosum*	0.019	64 (20/31)	67 (21/31)
15475	238	Conserved hypothetical protein	*Beggiatoa sp*	9e-21	67 (50/74)	77 (57/74)
15633	210	Thiamine monophosphate kinase (YP 528887.1)	*Saccharophagus degradans*	9e-19	52 (37/70)	74 (52/70)
16171	365	Endo 1,4 β-mannanase (YP 001982936.1)	*Cellvibrio japonicus*	6e-04	42 (24/56)	60 (34/56)

16256	323	B-1,4 Cellobioside	*Streptomyces sviceus*	8e-04	39 (29/74)	51 (38/74)
16362	135	B 1,4-Cellobiosidase (NP 821732.1)	*Streptomyces avermitilis*	0.12	54 (20/37)	59 (22/37)
16506	271	Endoglucanase (SP P37701.1)	*Clostridium josui*	2e-05	52 (18/34)	82 (28/34)
17088	508	Endo β-1,4 glucanase (ABI94085.1)	Uncultured bacterium	2e-40	46 (78/166)	66 (111/166)
17865	239	Endoglucanase like protein (YP 677947.1)	Cytophaga hutchinsonii	4e-16	55 (45/81)	67 (55/81)
18210	158	Thiamine monophosphate kinase (YP 528887.1)	*Saccharophagus degradans*	4e-11	60 (31/51)	74 (38/51)
18835	181	Endoglucanase like protein (YP 677948.1)	*Cytophaga hutchinsonii*	4e-07	43 (26/60)	65 (39/60)
19790	166	1,4 β-Cellobioside (YP 001902424.1)	*Xanthomonas campestris*	2e-06	47 (27/57)	64 (37/57)
20202	233	Hypothetical protein (XP658877.1)	*Aspergillus nidulans*	8e-04	62 (23/37)	72 (27/37)
20317	241	Chitinase (YP432187.1)	*Hahella chejuensis*	6e-10	40 (31/76)	60 (46/76)
20393	216	Beta-glucanase (ZP 01616464.1)	Marine γ-proteobacteria	0.004	38 (24/63)	55 (35/63)
20446	240	B-glucanase (EE 056766.1)	*Bacteroides sp*	7e-21	56 (45/79)	72 (57/79)
20483	241	Endo β 1,3-D glucosidase	*Shewanella piezotolerans*	3e-18	56 (45/79)	69 (55/79)
20900	103	Cellulase (BAB79288.1)	*Pseudomonas sp*	0.075	41 (14/34)	70 (24/34)
21307	241	Hypothetical protein (YP829062.1)	*Solibacter usitatus*	0.62	35 (21/59)	50 (30/59)
21541	241	Glycosyl transferase family 1 (ZP 01872231.1)	*Caminibacter mediatlanticus*	2e-18	57 (44/77)	70 (54/77)
21703	175	3-hydroxyacyl CoA dehydrogenase (YP 527046.1)	*Saccharophagus degradans*	3e-07	48 (25/52)	71 (37/52)

21717	104	Endo 1,4 β-glucanase (ZP 01113981.1)	*Reinekea sp*	8e-04	85 (18/21)	90 (19/21)
22358	241	Endo β 1,3-1,4 glucanase (YP 001615240.1)	*Sorangium cellulosum*	4e-16	55 (37/67)	74 (50/67)
22556	204	Hypothetical protein (NP106174.1)	*Mesorhizobium loti*	1e-16	69 (45/65)	81 (53/65)
22983	173	Endoglucanase like protein (YP 528465.1)	*Saccharophagus degradans*	1e-14	59 (34/57)	82 (47/57)
23140	241	Endoglucanase (YP435061.1)	*Hahella chejuensis*	2e-33	64 (51/79)	77 (61/79)
23287	241	Cellulase (GH5) (YP 002717928)	*Verrucomicrobiae bacterium*	8e-14	45 (35/77)	62 (48/77)
23614	224	1,4 β-Mannanase(ABY90130.1)	*Streptomyces sp.*	0.002	37 (23/61)	60 (37/61)
24116	123	Cellulase (BAB79288.1)	*Pseudomonas sp. ND137*	3e-04	52 (21/40)	72 (29/40)
24475	241	Hypothetical protein (ZP 01961605.1)	*Bacteroides caccae*	4e-16	50 (37/73)	71 (52/73)
25406	1025	Hypothetical protein (ZP 03852977.1)	*Chryseobacterium gleum*	6e-06	24 (34/141)	46 (65/141)
25464	207	Glycosyl Hydrolase family 5 (YP 002236921.1)	*Klebsiella pneumoniae*	1e-16	66 (44/66)	77 (51/66)
25876	222	Glycosyl hydrolase family 16 (YP 002508023.1)	*Halothermothrix orenii*	3e-19	53 (39/73)	73 (54/73)
25936	600	Glycosyl Hydrolase family 8 (YP 002715558.1)	*Verrucomivrobiae bacterium*	7e-45	48 (89/184)	64 (118/184)
26054	180	Cellobiohydrolase II (BAG48183.1)	*Irpex lacteus*	5e-08	50 (28/56)	62 (35/56)
26107	443	Bi-functional Xylanase/esterase (YP 677852)	*Cytophaga hutchinsonii*	2e-50	68 (98/144)	82 (119/144)
26243	176	2-isopropylmalate synthase (YP 527962.1)	*Saccharophagus degradans*	1e-13	62 (36/58)	77 (45/58)
26454	212	Cellulase (YP 001828440.1)	*Streptomyces griseus*	2e-08	43 (34/79)	50(40/79)

3.5 Discussion

Metagenomic sequencing projects are becoming an increasingly popular tool in microbial ecology, initiated by Craig Venter and colleagues (2004) with the shotgun sequencing of DNA extracted from Sargasso Seawater samples. That project produced 1.045 billion bp of non-redundant sequence, covering an estimated 1800 microbial species (based on sequence relatedness). 1412 16S rRNA genes or fragments of genes were recovered, leading to the designation of 148 previously unkown phylotypes. (at 97 % similarity cut off). Tyson *et al.* (2004) generated 76.2 million base pairs (bp) of DNA sequence data from an acid mine drainage system biofilm. In contrast to Venter *et al* (2004), the acid mine drainage biofilm was found to comprise a microbial communty of limited diversity including the recovery of two complete genomes and three nearly complete genomes. Biddle *et al.* (2007) generated 61.9 Mb of data in the form of 622, 129 reads with an average length of 100 bp deduced from DNA extracted from 4 sediment depth samples taken from Peru Margin, sub seafloor using the Roche 454 GS-20 pyrosequencer. Only 5-14 % (ranging depths) of reads produced had homology to sequences in the NCBI-nr database. Mou *et al.*, 2008, investigated a coastal marine assemblage (0.2-3 μM) metabolising DOC. Microcosms using sea water collected from Sapelo Island, Georgia were amended with the model DOC compounds dimethylsulphoniopropionate (DMSP) or Vanillate in the presence of the thymidine analogue bromodeoxyuridine (BrdU) and pyrosequencing of the newly synthesised DNA was performed. Bacterial assemblages without added DOC compounds were used as controls. Pyrosequencing of BrdU labelled DNA isolated by immunocapture produced 190,872 reads from duplicate DMSP-metagenomes and 115,807 reads from duplicate vanillate-metagenomes. 16S rRNA genes were shown to accout for 0.14 % of sequences. Taxa dominating both types of assemblage were γ-proteobacteria (61 % in DMSP and 53 % in vanillate). Primarily *Alteromonidales* and *Oceanospirillales* dominated. *Alphaproteobacteria* accounted for 16 % and 12 % of sequences and β-*Proteobacteria* accounted for 9 % and 11 % respectively. Brulc *et al.*, 2009 investiged the microbiome of bovine rumen by means of 454 pyrosequencing. More recently, a number of metagenomic sequencing projects from various sources have been

published (Warnecke *et al.*, 2007; Rusch *et al.*, 2007) with an increased interest in comparative metagenomic studies whereby multiple metagenomic datasets are compared against each other (Tringe *et al.*, 2005; Huson *et al.*, 2009; Qi *et al.*, 2009). A number of gene annotation methods are available (for review, see Kunin *et al.*, 2008). The GH repertoire of bovine rumens (known to contain complex microbial communities for the degradation of plant polysaccharide material) have previously been investigated using 454 pyrosequencing technology, where Brulc *et al.*, (2009) sequenced the four metagenomes (three of fibre adherent microbiomes from different bovine animals fed the same diet and one pooled liquid biome of the same animals). Collectively the sequencing provided ca. 1 million sequence reads resulting in 103,929,743 bp. The communities were dominated by bacteria when comparing the pyrosequencing data to SEED subsystems and the RDP database. When pyrosequencing reads were compared to all catalytic and non-catalytic modules generated from the CAZy databse, a total of >3,800 GH were matched belonging to 35 GH families and nine CBMs belonging to three families.

With the inherent amount of data generated from metagenomes manual curation of resultant genetic material is not feasible. Similarly data management of resulting matches also requires automated procedures. The method used here for prediction of metagenomic material was mostly automated enlisting BLAST for taxonomical and functional comparison and assignment of the Irish Sea metagenome sequences against the NCBI non-redundant protein and SEED database, coupled with MEGAN and MG-RAST data management programs respectively. Similar taxonomical relationships were seen for both functional gene and 16S rDNA analysis with Proteobacteria and *Bacteroidetes* dominating. Therefore taxonomical assignment of randomly pyrosequenced community DNA enables an initial insight into the population composition of a dynamic and complex bacterial biofilm.

Further insights into the community were possible through analysis of the glycosyl hydrolase (GH) gene repertoir of the population. Most of knowledge regarding GH's in the marine environment has originated from studies on cultured species, such as *Saccharophagus degradans* (Taylor *et al.*, 2006) which can degrade several

polysaccharides. A few molecular culture - independent studies have focused on GHs in the marine environment, mainly focusing on chitinases as chitinase is known to be widespread amongst readily cultured members of the *Vibrio* group such as *V. Alginolyticus* (Ohishi *et al.,* 2000). Cottrell *et al.* (1999), screened metagenomic libraries derived from environmental DNA from coastal and esturine water using fluorogenic analogues of chitin and cellulose identifying a number of genes involved in chitin hydrolysis however no genes were identified from screening with the cellulose analogue. The diversity and abundance of β-1,4 endo glucanases within the GH5 family in the North Altantic Ocean has been investigated by designing primers for a 350 bp fragment of eight β-1,4 endoglucanases belonging to the GH5 family, constituting ~1/3 of the gene. Relative abundance was determined for 3 locations using qPCR where abundance was found to positively correlate with chlorophyll concentrations. Using clone libraries the GH5 family genes were found to be more diverse in coastal water than open ocean (Elifantz *et al.,* 2008). Cottrell *et al* (2005) employed screening of a fosmid library consructed of prokaryotic DNA from the Western Arctic Ocean. Primers were designed for the most abundant type of endoglucanase identified in the Sargasso Sea WGS data set (Venter *et al.,* 2004). However, subsequent biochemical protein analysis revealed the gene to encode a peptidase. The research reported in this thesis is the first study designed to specifically assess the bacterial communities that colonise cellulose in the marine environment and to concomitantly evaluate the glycosyl hydrolase (cellulase and chitinase) gene repertoire of that communty, in the absence of the biases associated with PCR based molecular techniques.

Although 116 contigs resulted in matches against the assembled GH database it can not be assumed that all the contigs encode enzymes with cellulase or chitinase activity. The diversity of polysaccharide structures in nature is matched by an equally diverse array of enzymes responsible for their degradation, and although the GH families were selected on the basis that they contained representatives of cellulases and chitinases, some (i.e. GH5, GH 8, GH12, GH16) encompass enzymes with activity against other substrates. Analysis based on sequence similarities of DNA from environmental sources has previously been shown to be misleading; for example

Cottrell *et al.*, (2005) designated a gene from an uncultured bacterium as likely to encode a cellulase, but it was later shown to express protease activity.

Chitinase genes have been found in α- and *γ-proteobacteria* (Cottrell *et al.*, 2000) and the *Cytophaga-Flavobacteria-Bacteroides* group (Xaio *et al.*, 2005) and are thought to be widely distributed in marine bacteria. Both families known to represent chitinases (GH18 and 19) were included in the GH database and although representatives of these groups of bacteria are present in the Irish Sea dataset at abundant levels, only a small number of contigs gave a top hit to a chitinase (contigs 06951, 10324, 11489, 20317), suggesting that a functionally specialised consortia of bacteria, involved in the cellulose degradation process are specifically colonising the cellulose bait. Further work to investigate this hypothesis would be the random pyrosequencing of DNA collected from a biofilm community attached to a chitin bait of the same location.

Heterotrophic marine bacteria play an important role in organic carbon recycling through degradation of POM and DOM, of which complex polysaccharides such as cellulose and chitin are major constituents (see chapter 1). It has been shown through the analysis of 16S rDNA that population structures differ between free living and particle attached communities. DeLong *et al.*, 1993 found that *Alphaproteobacteria* dominated free living bacterioplankton whilst *Cytophaga, Gammaproteobacteria* and *Planctomyces* were abundant when looking at particle attached bacterial communities. However rRNA gene based assessments are prone to bias either through the PCR reaction itself (Osborn &Smith, 2005) and through primer sets used as they are based on reference sequences of the 'known fraction'. Metagenomic studies provide an alternative mechanism for environmental microbial population analysis, either through protein or rRNA gene sequence similarity or random sequencing of the total community DNA excluding PCR bias.

The most challenging obstacle to overcome in metagenome annotation is the limited microbial representation of most environments in reference databases, constituting mainly cultured isolates. Additionally short read length, inherent genetic heterogeneity and inter-species gene conservation are barriers to the assembly of

environmental 454 sequence reads into contiguous sequences (contigs) (Krause *et al.*, 2008), add to the challenges faced in this field. However new problems with correct annotation is likely to increase with the increase of computer based genome and metagenome sequencing projects. There is clearly a fine balance in metagenomic analysis between accurate annotation and being able to recognise potential protein encoding genes from unknown microorganisms. There are currently 826 (May 2009) genomes sequenced on the NCBI database of which 213 (26%) belong to the *Gammaproteobacteria*, 105 (12 %) to the *Alphaproteobacteria* and 26 (3%) to the *Bacteroidetes.* Thus the most abundant sequenced genomes will match the majority of sequence data leading to a skewing of the results. However 184 (22 %) of sequenced genomes are from the *Firmicutes* but only represented by 73 assigned contigs using MEGAN analysis and 1.87 % of assignments using MG-RAST analysis. It is well established that Gram negative bacteria, and *Gammaproteobacteria* in particular, predominate in the marine environment, and Gram positive bacteria are less abundant. It is however obvious that assignments made by these programs are simply closest known matches and not evidence for the presence of a particular species. There are currently more genome sequencing projects in progress than are represented in the database and only with more sequencing of the culturable proportion and the unculturable proportion (single cell genomics) (Woyke *et al.*, 2009) will this problem be circumvented. There is no doubt however that this is now occurring at a fast rate and re-analysis this dataset in ca. 12 months to compare the MEGAN and MG RAST assignments could well produce significantly different results.

3.6 Conclusions

- Community DNA was extracted from cellulose bait colonising microorganisms and subjected to random 454 pyrosequencing.
- Assembled contigs were analysed on the basis of taxonomy and function using a number of bio-informatical methods
- A blastx comparison against the NCBI-nr protein database resulted in a large number of matching sequences being taxonomically assigned and showed members of the *Proteobacteria* and *Bacteroidetes* to predominate. Additionally functional assignment based on the NCBI microbial attributed table was performed showing a high percentage of genes matched to sequences in the reference database indicated as involved in the carbohydrate metabolism process.
- Further taxonomic assessment using the MG-RAST data management website service supported the evidence that *Proteobacteria* and *Bacteroidetes* are dominant in the biofim community, via functional gene analysis and analysis of the 16S ribosomal DNA phylogenetic marker.
- A wide repertoire of genes were located by comparison with a limited customised glycosyl hydrolase database.

Chapter 4

Metaproteomic analysis of biofilm communities colonising cellulose and chitin baits in the Irish Sea

4.1 Introduction

4.1.1 Metaproteomics

Proteomics is a functional application which identifies and characterises the total set of proteins from a cell or organism that collectively are known as the proteome (Phillips & Bogyo, 2005; Li, *et al*, 2004). Studying the proteome is complex and challenging, as in any given organism there is a wide array of proteins that may be differentially expressed, post-translationally modified or act only when in the form of complexes (Gavin *et al*, 2002). Owing to the high complexity of cellular proteomes and low abundance of many of the proteins expressed, proteome based studies pose significant challenges (Aebersold & Mann, 2003). Metaproteomics is an order of magnitude more complex than proteomics and was first defined by Wilmes & Bond (2004) as involving the analysis of the proteome of an entire microbial community at a given point in time and referred to as a metaproteome.

The majority of research concerning microbial communties has been at the genetic level and until recently has focused mainly on phylogenetic diversity. Now with the growing number of metagenomic sequencing projects (e.g. Venter *et al.*, 2004) it is becoming evident that more information about complex microbial communtiy function is required. The presence of a gene or a pathway does not necessarily lead to its expression or expected function. Metaproteomics has the potential to reveal functional information on natural microbial communties, providing an insight into key biochemical activities that occur, along with responses to stresses such as temperature, osmotic pressure and bacterial interactions within a natural assemblage.

Proteomics is a recently established discipline and already plays a significant role in characterisation of bacteria under pure culture conditions, for which genomic data is also available (Gade *et al.*, 2005; Ekborg *et al.*, 2006; Oda *et al.*, 2006). Although advantageous, pure culture proteomics inherently has limitations in environmental

microbiology, notably the paucity of bacterial species that can be cultured under standard laboratory conditions, and consequently the inability to study the proteins that they express under variable environmental conditions.

Probably the most successful metaproteomic study to date was that of Ram *et al.* (2005) in which 'shotgun' MS based proteomics was applied to provide a high-throughput method to positively detect 2033 proteins expressed by a low complexity biofilm community known to contain 5 dominant species: *Leptospirillum* group II and group III, *Sulfobacillus* and Archaea related to *Ferroplasma acidarmanus* and "G-plasma". Two dimensional PAGE based proteomics and later a shotgun approach have been applied to a metaproteome from a laboratory scale activated sludge treatment system to study enhanced biological phosphorous removal (EBPR) (Wilme & Bond, 2004; Wimes *et al.*, 2008). Powell *et al.* (2005) employed 1D SDS-PAGE and MS to identify a small number of proteins recovered from the dissolved protein in seawater. Klaassens *et al.* (2007) combined 2D PAGE with MALDI-TOF MS to analyse a limited microbial community of human faecal samples from new born babies, leading to the identification of a transaldolase protein from *Bifidobacterium infantis*, a known species of the microbiota in the human infant gastrointestinal tract. Verberkmoes *et al.* (2009) applied shotgun proteomics to investigate the functional diversity of human faecal samples; differentially expressed proteins from bacterial communities following exposure to cadmium have been investigated (Lacerda *et al.*, 2007) as well as contaminated soil and groundwater (Benndorf *et al.*, 2007). Strain-resolved community proteomics of an acid mine drainage system and activated sludge have recently been implemented by Lo *et al.* (2007) and Wilmes *et al.*, (2008). Lo *et al.* (2007) extracted total protein from a genomically uncharacterised biofilm of the 'ABend location' within the Richmond mine, California, USA and compared generated peptide mass spectra to two community genomes generated from a '5-way CG' Richmond mine metagenomic dataset and 'UBA location' Richmond mine metagenomic dataset. The two community generated datasets were dominated by bacteria belonging to *Leptospirillum group* II, however composite strains showed variation between the two biofilm samples. From a third genomically uncharacterised 'ABend location' within the Richmond mine, protein

extracted resulted in 7,526 peptides matching *Leptospirillum* group II proteins from a '5-way CG' Richmond mine metagenomic dataset but not the 'UBA location' *Leptospirillum* group II proteins and 23,787 peptides matched a 'UBA location' Richmond mine metagenome and not the a '5-way CG' Richmond mine *Leptospirillum* group II protein variants.

4.1.2 Protein extraction

Preparation of protein extracts is probably the most critical step in proteomic projects, as extraction techniques are targeted towards specific protein localisation (extracellular, membrane associated and intracellular). Differential expression and the complexity of the cellular localisation of proteins ultimately results in a dynamic range of extraction techniques, particularly when dealing with complex microbial communities. A number of cell disruption methods are widely used, from gentle (enzymatic digestion) to more stringent methods (French press, sonication, bead beating); if extracellular proteins are targeted then lysis of cells is unnecessary. Fractionation of proteins can therefore be achieved to some extent by the extraction method chosen. Previous metaproteomic projects have primarily focused on collection of cells and subsequent lysis. Wilmes & Bond (2004) combined centrifugation with washing of bacterial cells from activated sludge followed by lysis using a French Press; Kan *et al.* (2005) combined tangential flow concentration with centrifugation for isolation of planktonic microbial cells, from Chesapeake Bay, and lysis using a Fench press method; Pierre-Alain *et al.*, (2006) isolated bacterial cells using high speed density gradient centrifugation from freshwater samples, whilst Verbekmoes *et al.* (2009) also utilised differential centrifugation to collect bacterial cells from human faecal samples which thereafter were lysed with guanidine and dithiothreitol. Lo *et al.* (2007) fractionated proteins into extracellular, membrane, soluble and whole cell from an acid mine drainage biofilm.

4.1.3 Liquid and Gel based Protein Separation

Chromatography is used in proteomics for the preparative separation and purification of mixtures of proteins and peptides based on a number of their features. Reverse phase chromatography separates proteins and peptides based on their reversible interaction with the hydrophobic surface of the column matrix and is often the final stage of separation prior to sequencing of peptides and proteins. In hydrophobic interaction chromatography, proteins are loaded in the presence of a high salt concentration and eluted with a decreasing salt gradient, separating proteins based on their surface hydrophobic character. Separation of proteins based on their charge is performed using Ion Exchange Chromatography (IEX) - cation exchange chromatography (CatIEX) or anion exchange chromatography (AnIEX), whilst separation of proteins based on size is conducted by size exclusion chromatography (SEC) also known as gel filtration chromatography.

Polyacrylamide gel electrophoresis is a commonly used gel based technique for separation and analysis of protein mixtures and is most frequently used in combination with sodium dodecyl sulphate (SDS; Shapiro *et al.*, 1967). The anionic detergent SDS binds to proteins resulting in a net negative charge (~1.4 g of SDS per gram of polypeptide; Reynolds & Tanford, 1970). As part of the Laemmli (1970) one-dimensional polyacrylamide gel electrophoresis (1D SDS-PAGE), a Tris-glycine SDS discontinuous buffer system ensures that proteins are dissociated, unfolded into polypeptide chains, and migrate through the polyacrylamide matrix under electrophoretic conditions. Such dissociation of proteins enables separation based solely on the molecular weight which can be determined to within 10 % of the true value when compared against a known standard (Simpson, 2003). However a number of other systems can be applied including non-denaturing or native PAGE (for analysis of native proteins and protein complexes), separating proteins based on size and charge properties as well as continuous buffer systems and the development of gradient gels (Simpson, 2003).

Further separation of proteins can be achieved with two-dimensional (2D) PAGE. 2D PAGE is a powerful and widely used method for the analysis of complex mixtures of proteins and is a central technique in proteomics (Celis & Gromov, 1999).

The method was first developed by O'Farrell (1975), with modifications made by Bjellqvist *et al.* (1982) replacing carrier ampholyte-generated pH gradients with immobilised pH gradients (IPG). According to the method of O'Farrell (1975), proteins are separated by two parameters; firstly on the basis of their net charge or their isoelectric point (pI) in an isoelectric focusing (IEF) step; secondly, according to their molecular weights on a denaturing polyacrylamide gel (SDS-PAGE). This gives the ability to separate or resolve a much larger number of proteins.

4.1.4 Mass Spectrometry

A number of mass spectrometer platforms are available for MS based proteomics and all consist of three main components: an ion source, a mass analyser, and an ion detector. Ionisation converts molecules to ions so they can be manipulated within electric or magnetic fields without degradation or fragmentation (Yates, 2004). Electrospray ionisation (ESI) ionises the analytes out of a solution (Fenn *et al.*, 1989) and is readily coupled with liquid based separation tools; Matrix Assisted Laser Desorption Ionisation (MALDI), ionises samples out of a dry crystalline matrix via laser pulses. Both are 'soft ionisation' techniques so as not to cause fragmentation. Analysers include Ion trap, Time of Flight (TOF), quadrupole (Q) and Fourier transform ion cyclotron (FT-MS), with key parameters being sensitivity, resolution, mass accuracy and the ability to generate information- rich ion mass spectra from peptide fragments (MS/MS spectra). No single instrument or method is capable of identifying and quantifying the components of a complex protein sample in a single step.

Initially, identification of proteins was through mass measurement of enzymatically digested peptides. Peptide mass fingerprints (PMF) retrieved from such experiments can be compared to predicted masses of known proteins in databases. PMF-MS is commonly combined with 2D gel based protein separation. However results do not meet the specificity required for protein mixtures and for proteins for which there are no reference data, e.g. uncultured bacterial species. Tandem mass spectrometry (MS/MS or MS^2) can provide PMF data or can be used for *de novo* peptide sequencing for proteins present in mixtures at low femto mole levels

(McCormack *et al.*, 1997). This involves selection of a parent ion in the first mass spectrometer, fragmentation by a collision event by collision-induced dissociation (CID), surface-induced dissociation (SID) or photon-induced dissociation (PID) and the m/z values for the resulting daughter ions measured in a second mass spectrometer, resulting in a series of ions which can contain sufficient information to determine a peptide sequence (Standing, 2003). Combinations include: TOF/TOF, Q/TOF and Linear Quadrupole Ion Trap (Yates, 2004). MS/MS can be applied to isolated proteins which have been enzymatically digested or for peptide mixtures using a 'shotgun' approach (Delahunt & Yates, 2005) in which protein mixtures are enzymatically digested and the resultant peptides separated by online chromatography, commonly using reverse phase liquid chromatography (RP-LC) for 1D separation or combining strong cation exchange chromatography (SCX) with RP-LC for 2D separation in a technique known as LC/MS/MS and 2D-LC/MS/MS. (Delahunty & Yates, 2005). However affinity and size exclusion chromatography can also be used (Delahunty & Yates, 2005). Recent advances in capture efficiency and storage capacity have provided rapid and sensitive MS/MS/MS, or MS^3 using linear ion trap mass spectrometers (Hager, 2002). Approximately three fold enhancement in data depth has been achieved with the new generation linear trapping quadrupole technology, such as the LTQ-FT-Orbitrap (Quadrupole ion trap- Fourier transform ion cyclotron) hybrid instrument (Banfield *et al.*, 2005), with the introduction of an additional round of fragmentation coupled with advances in computation algorithms to score MS^3 spectra (Olsen & Mann, 2004). This will provide a means of peptide sequencing of complex mixtures of proteins even at subfemtomol levels, and with high mass accuracy (Banfield *et al.*, 2005).

4.1.5 Aims and Objectives

The aims of this chapter are firstly to develop a method for the extraction of the extracellular protein fraction from the biofilm community colonising cotton cellulose and crab shell chitin bait recovered from the Irish Sea. Secondly, recovered proteins will be analysed for cellulase and chitinase activity by means of zymograms. Finally, following successful extraction and purification of community proteins Mass Spectrometry will be applied to the identification of proteins predicted to be involved in cellulolysis and chitinolysis.

4.2 Methods

4.2.1 Protein Extraction

For more detail see page 126, briefly cellulose bait material, moored in the Irish Sea (as described chapter 2) was weighed and known quantities of cellulose bait were placed into a 50 mL screw cap centrifuge tube with half the weight of bait in volume of NaCl (4 M). The bait was gently shaken at 4°C for 1 h. Any excess liquid was collected in a fresh 50 mL screw cap centrifuge tube and residual liquid retained by the bait was extracted (see figure 4.1).

4.2.2 Chromatography of community extracted proteins

4.2.2.1 Anion exchange chromatography (AnIEX)

The Proteins extracted from cellulose bait returned from the Irish Sea (see chapter 2) using 4 M NaCl were dialysed against 50 mM Tris-HCl, pH 8 for the removal of salt. The dialysed protein sample was applied to a Hi Trap Q HP column (GE Healthcare) equilibrated with buffer A (50 mM Tris-HCl, pH 8) at 0.5 mL min^{-1}. Protein was collected with a gradient of 0 - 1 M NaCl. Elution was monitored by continuous measurement of the absorbance at 280 nm. Fractions were analysed for endoglucanase activity by zymograms.

4.2.2.2 Hydrophobic interaction chromatography

The fractions found to possess endoglucanse activity following zymogram analysis were fractionated by hydrophobic interaction chromatography. The dialysed protein solution was added to 3 M NaCl and applied to a phenyl sepharose CL-4B column (GE Healthcare) also equilibrated with 3 M NaCl. Proteins were eluted with 20 mM phosphate buffer, pH 7.0 in fractions of 1 ml and the elution profile obtained by continuous measurement of the absorbance (280 nm).

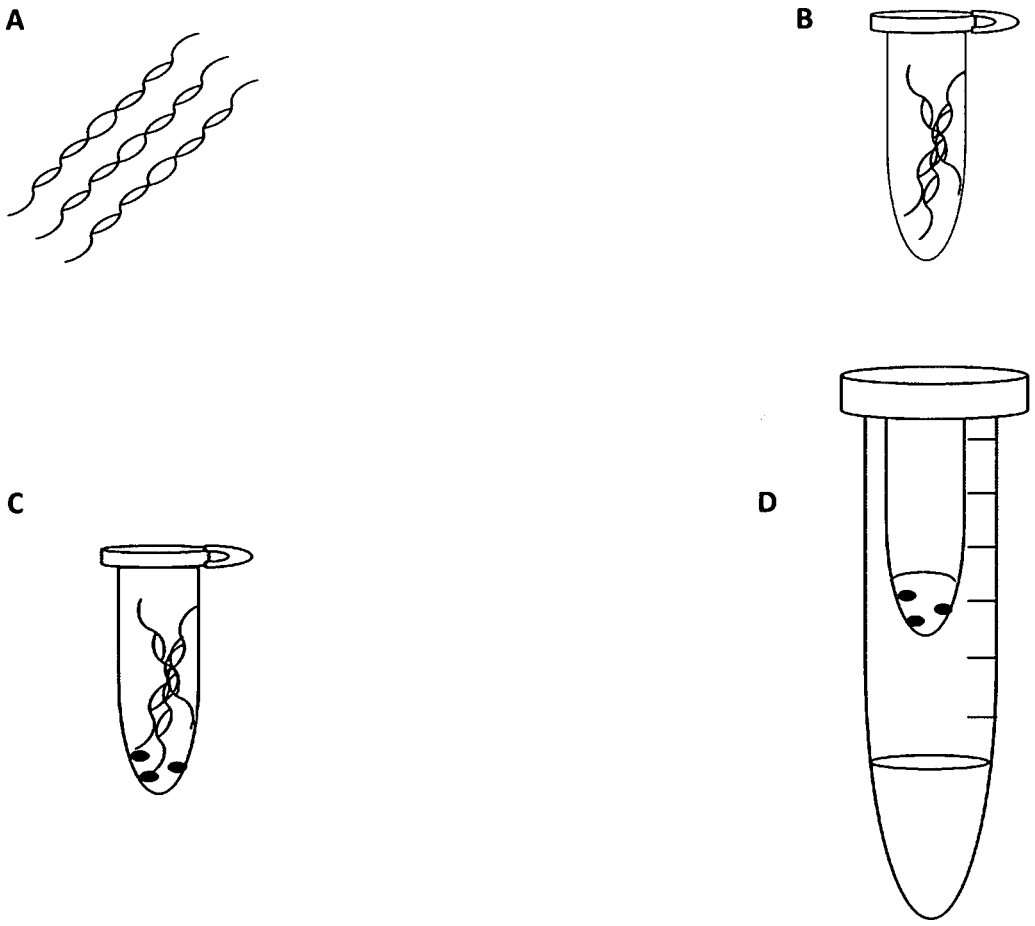


Figure 4.1 Liquid Extraction from cellulose bait

Cellulose string was cut into pieces of approximately 2 cm with sterile scissors **(A)** and packed into 1.5 mL microfuge tubes using sterile forceps **(B)**. Holes were made in the base of the microfuge tube with a heated needle from a dissection kit **(C)** and the microfuge tube placed into a screw cap 15 mL centrifuge tube, which was centrifuged at 5000 x g for 10 min to extract all liquid from the string bait material **(D)**.

4.2.3 Polyacrylamide gel electrophoresis (PAGE)

4.2.3.1 Zymography

Ethylene glycol chitin (EGC) (Seikagaku) for detection of chitinase or CMC for endoglucanase activity was incorporated into the separating portion of an SDS-PAGE gel to a final concentration of 0.01 % or 0.1 % respectively. Protein extract was added to loading buffer (125 mM Tris-HCl, 15 % (w/v) sucrose, 2.5 % SDS and 0.02 % (w/v) bromophenol blue) (Pedra Reyes & Gutierrez-Corona, 1997) and heated at 80°C for 10 min. Following fractionation of proteins, the zymogram was washed 2 x 20 min in ddH_2O and incubated in 0.1 M sodium phosphate buffer, 1 % Triton X-100, pH 6.4 overnight at 37°C and subsequently analysed for chitinase or cellulase activity by staining zymograms with 0.1 % Congo red (Sigma) for 30 min and destaining with 2 x 15 min with 1 M NaCl.

4.2.3.2 One-Dimensional SDS-PAGE

Polyacrylamide gel electrophoresis was carried out as described by Laemmli (1970) using 10 % (w/v) polyacrylamide resolving gels and 4 % (w/v) polyacrylamide stacking gels, unless otherwise indicated. Samples were run at 90 V through the stacking gel, and then 120 V through the resolving gel. Gels were stained with Coomassie Blue R-250 and destained with 10 % (v/v) methanol: 10 % (v/v) acetic acid. A broad range protein marker (Bio-Rad) was used unless otherwise indicated.

4.2.3.3 Two Dimensional Gel Electrophoresis

Extracted proteins were added to rehydration buffer (8 M urea 2 M thiourea and 4 % CHAPS with the addition prior to use of 3.4 mg DTT, 2 µl ampholytes (Bio-Rad) and 10 % ASB14) to a final volume of 135 µl. Protein/rehydration solution was left at room temperature for 1 hour with vortexing every 15 min, then centrifuged at 8000 x g for 5 min and 125 µl applied to a 7 cm non linear IPG strip (Bio-Rad) pH 3-10. Following focusing, the IPG strips were equilibrated and applied to mini gels containing a 10 % resolving and 4% stacking gel and sealed with agarose (1 %) and bromophenol blue.

Second dimension separation was carried out at 130 V. Gels were stained with colloidal Coomassie (Severn Biotech) as per manufacturer's instructions or with a double silver stain (Heukshoven & Dernick, 1985).

4.2.3.4 Coomassie stain

100 mL Coomassie stain (25 % isopropanol, 10 % glacial acetic acid, 0.25 % Coomassie R-250 (Sigma)) was added per polyacrylamide gel and incubated at ambient temperature for 1 h. Coomassie stain was removed and gels were destained by the addition of 10 % methanol; 10 % acetic acid made up in dH_2O.

4.2.3.5 Silver stain (Heukshoven & Dernick, 1985)

Polyacrylamide gels were fixed for 1 h in fixing solution (40 % ethanol, 10 % acetic acid, 50 % distilled water). The fixing solution was discarded and gels washed twice in 10 % ethanol for 5 min, followed by two 5 min washes with distilled water. The washed gel was incubated, shaking for 30 min in $AgNO_3$ (0.2 % in dH_2O) at ambient temperature after which the gel was washed for 5 sec in dH_2O and the gel developed to the required intensity by the addition of developing solution (2.5 % sodium carbonate, 0.08 % formaldehyde in dH_2O). The reaction was terminated by the addition of acetic acid (1 %), after which the gel was washed in dH_2O for 6 X 8 min. The gel was reduced until clear with the addition of sensitising solution (0.05 % sodium carbonate, 0.15 % potassium ferricyanide, 0.3 % sodium thiosulphate in dH_2O). Gels were washed 4 X 5 min with dH_2O and $AgNO_3$ added and again incubated for 30 min. Gels were washed in dH_2O for 5 secs and developing solution added until the intensity required was reached, at which point the reaction was terminated with the addition of 1 % acetic acid and washed in distilled water 6 X 8 min. If a high background staining persisted, gels were again submersed in sensitising solution, followed by washing 4 X 5 min in dH_2O.

4.2.4 Phenol extraction

An equal volume of phenol (pH 7.6) (Sigma) was added to extracted protein. The mixture was vortexed and heated for 10 min at 70 °C and cooled on ice for 5 min. The upper aqueous layer was then removed and an equal volume of dH_2O was added. The mixture was again heated for 10 min at 70 °C and cooled on ice for 5 min and centrifuged at 5000 x g for 10 min at 4 °C. The upper layer was removed and 2x volumes of ice cold acetone added, and protein precipitated overnight.

4.2.5 Trypsin digestion of proteins

4.2.5.1 In solution trypsin digestion

500 µg protein (determined using Sigma's BCA method according to manufacturer's instructions) was pelleted by centrifugation at 1000 x g under vacuum. The pellet was resuspended in 100 µL urea (8 M) and ammonium bicarbonate (0.4 M) to which 10 µL dithiothreitol (DTT) (50mM) was also added and then incubated for 15 min at 50°C. Following incubation, 10 µL 0.1 M iodoacetamide (IAA) was added and incubation continued for a further 15 min at room temperature in the dark. The solution was then made up to a volume of 400 µL with ddH_2O. To the remaining solution, 15 µg Gold Mass Spectrometry grade trypsin (Promega) was added and incubated overnight at 25°C. A further 7.5 µg trypsin was then added and the solution incubated for a further 1 h.

4.2.5.2 In gel trypsin digestion

Excised protein bands were washed twice with wash buffer (50 % acetonitrile; 200 mM ammonium bicarbonate). Wash buffer was discarded and the gel slice allowed to air dry for one hour. 5µL RB buffer (0.2 M ammonium bicarbonate, 2 M urea) containing 0.1 µg trypsin Gold Mass Spectrometry grade (Promega) was added and allowed to rehydrate the gel slice. The gel slice was broken into 4 pieces and 15 µL RB buffer added. The gel slice was incubated overnight at room temperature. Excess

buffer was removed to a clean tube and 20 µL extraction buffer was added to the gel slice (60% acetonitrile, 0.1% trifluoroacetic acid (TFA)), and the tube gently shaken for 30 min. Excess buffer was removed and pooled with that previously removed and the process repeated. The extracted peptides were centrifuged under vacuum to dryness, 50 µL water added and again centrifuged under vacuum to dryness. Peptides were then purified using a Zip Tip pipette tip (Millipore) according to manufacturer's instructions.

4.2.6 Two dimensional liquid chromatography (2DLC)

Enzymatically digested peptides were desalted by loading onto a 5 mL Sephadex G25 HiTrap column (GE Healthcare) and eluted with potassium phosphate (10 mM) pH 2.7 containing 20 % Acetonitrile. Desalted peptides were collected and dried by centrifugal evaporation, and resuspended in 100 µL buffer A (potassium phosphate (10 mM) pH 2.7 containing 20 % Acetonitrile).

4.2.6.1 Cation exchange chromatography (CatIEX)

Peptides were applied to a PolySulfoethyl A™ 200 Å200, 5 µM (50 x 4.0 mm) column (Poly LC, Columbia, USA) which was equilibrated with buffer A. Peptides were separated with a gradient of 0 - 0.25 M KCl in buffer A for 40 min then 0.25 - 0.6 M KCl in buffer A for 15 min with a flow rate of 0.25 mL min^{-1} and elution was monitored at 280 nm.

4.2.6.2 Reverse phase chromatography (RPC)

Peptides were collected and concentrated by centrifugal evaporation and applied to a PepMap C18 Reverse Phase-HPLC column (100 x 2.1 mm) (Applied Biosystems) which had been equilibrated with 0.1 % TFA. Peptides were separated with a 30 min gradient of 0-64 % acetonitrile in 0.1 % TFA. Elution was monitored at 214 nm.

4.2.7 Mass Spectrometry and peptide sequencing

Peptides were applied to an UltiMate nano-liquid chromatography column (LC Packings) connected to a Waters (Manchester, UK) Q-TOF Micro electrospray tandem mass spectrometer, operated in positive ion mode. Chromatography was performed on a μ-Precolumn C18 cartridge (LC Packings) connected to a PepMap C18 column (3 μm 100 Å packing; 15 cm x 75 μm i.d.), using a linear gradient of 5% (v/v) solvent B [0.1% v/v formic acid in 80% (v/v) acetonitrile in water] in solvent A [0.1% (v/v) formic acid in 2% (v/v) acetonitrile in water] to 100% solvent B over 60min at a flow rate of 200nl/min. The spectrometer was operated in Data Directed Analysis (DDA) mode, where a survey scan was acquired from m/ζ 400-1500, with switching to MS/MS on multiply charged ions. MS/MS mass spectra were acquired over mass range 80-2000 Da. Partial peptide sequences were determined by manual interpretation of MS/MS data using the PepSeq software within the MassLynx package (Waters).

4.3 Results

4.3.1 Sampling

Following one month moored *in situ* in the Irish Sea, cellulose and chitin baits were returned to the laboratory, catalogued and frozen at -80°C until required.

The sampling of cellulose string was more effective than that for chitin, very little of which remained intact after one month at sea (Figure 4.2). A number of possible reasons may account for this. Firstly, the physical nature of the crab shell chitin when subjected to the buoyant environment probably caused physical erosion of adjacent pieces reducing the surface area for biofilm development. Secondly, chitin is almost certainly degraded faster than highly crystalline cotton cellulose. Furthermore, bags containing chitin regularly returned empty, probably due also to the abrasive action of the chitin against the nylon bags causing holes to form and loss of the bait. Bags of cellulose baits rarely returned with damage and the string was always retained. It would have been desirable to investigate decreasing the residence time of the chitin baits but due to the experimental cruise regime employed by the Proudman Oceanography Laboratory, this was not possible.

4.3.2 Development and optimisation of methods for extraction of proteins from cellulose bait

Initial extraction methods were applied to cellulose baits only because the amount of starting material was much greater. There is no universal method for extracting proteins and the approaches used depend on the localisation of protein (extracellular, membrane-bound or intracellular) and whether proteins are required in denatured or native form. This is further complicated for proteins to be extracted from a complex biofilm. Initial methods were developed empirically by 'trial and error' and included boiling string in denaturing laemmli buffer (Laemmli, 1970) and ribolysation (bead beating) of string. Neither of these resulted in a significant yield of protein when examined by SDS-PAGE. Cellulases and chitinases are mostly excreted by bacterial cells into the surrounding area or are bound to the surface of the cell. Requirement for a more targeted approach to collect such proteins whilst limiting disruption of bacterial cells was required to minimise the heterogeneity of the extracted protein sample.

The issue of extracting proteins without significant cell disruption was approached by using zymograms to enable visualisation of active cellulases and chitinases in a protein extract. The first approach was gentle agitation of the colonised cellulose in buffer for one hour at 4 °C. A number of buffers were used, including UTUCHAPS (7 M urea; 2 M thiourea; 2% CHAPS; 1 % DTT; 1 complete EDTA free protease inhibitor tablet (Roche) per 10 ml buffer) and UTUCHAPS/CTAB (CTAB introduced in an attempt to minimise coextraction of carbohydrates) (7 M urea; 2 M thiourea; 1% CHAPS; 1 % CTAB; 1 % DTT; 1 protease inhibitor tablet per 10 ml buffer). Proteins were successfully extracted from cellulose bait albeit with significant smearing and background staining of the SDS-PAGE gels and endoglucanase activity was visualised for a number of protein bands by zymogram analysis of proteins extracted with a UTUCHAPS/CTAB buffer (figure 4.3). It was postulated that the smearing and streaking was caused by co-extraction of lipids, polysaccharides and other polymers released by microorganisms into the biofilm matirix. Although early work was performed using UTUCHAPS/CTAB buffers, the probability of cell lysis resulting from the strong denaturing properties of the buffer would unnecessarily increase sample complexity.

It was reasoned that NaCl would release interactions of protein with the substrate, whilst agitation would release protein attached to cells and from the biofilm matrix. When comparing extraction of protein with either UTUCHAPS or 1 M NaCl by gentle agitation, it was observed that there was increased protein retrieval when extraction buffer was added as compared to simple removal of residual liquid from the string, following gentle agitation for 1 h. In addition, early approaches included a number of protein precipitation methods (including TCA, and TCA in acetone). Figure 4.4 shows the increase in protein retrieved with the precipitation of protein using TCA in acetone when compared to TCA alone. Although precipitation methods improved the recovery of proteins, there were subsequent handling problems, it also increased the level of contaminating products. Protein pellets were characterised by a waxy texture and were difficult to solubilise and clearly contained material other than just protein. This resulted in difficulties with protein quantification, for which Bradford,

BCA (Sigma) and 2-D Quant kit (Amersham Biosciences) were all tested but proved unsuccessful. This meant that method assessment was difficult, other than visualisation by 1D SDS-PAGE. Precipitation methods rely on the presence of sufficient protein to cause precipitation and often very little protein would precipitate due to the small amounts of starting material available.

When protein extracts performed with UTUCHAPS, UTUCHAPS/CTAB and NaCl were visualised by zymograms, bands that were absent with other buffers (UTUCHAPs and UTUCHAPs/CTAB) were visible with NaCl extracted proteins, when an extraction was made from the same bait source with loading of identical volumes of extract (figure 4.5). This was investigated further and a gradient of sodium chloride was added to pieces of cellulose bait of the same weight in 50 mL screw cap tubes at increasing concentrations (0 M- 4 M) and gently shaken for one hour. Noteably, the volume of liquid removed from the sample was double that added due to the retention of seawater by the cotton string, effectively reducing the concentration by ca. 50%. All liquid was removed from the string (Figure 4.1) and analysed by zymogram (figure 4.5). These data show that the amount of activity (intensity of clearing on zymogram) is proportional to the NaCl concentration used to extract proteins from the bait.

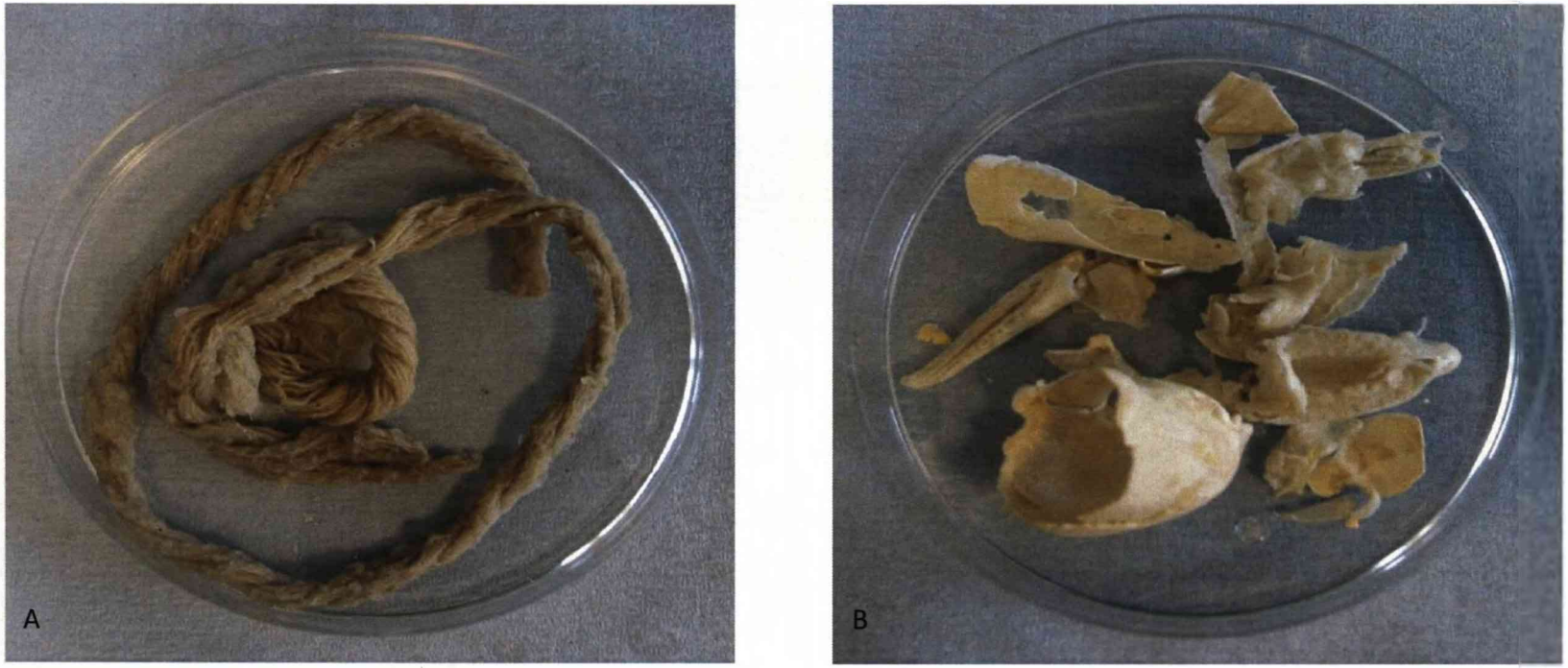


Figure 4.2 Cellulose and chitin baits used for microbial community analysis

Cotton cellulose **(A)** and crab shell chitin **(B)** were moored at two sites in the Irish Sea for one month. Baits were used as material for microbial colonisation and community microbial analysis.

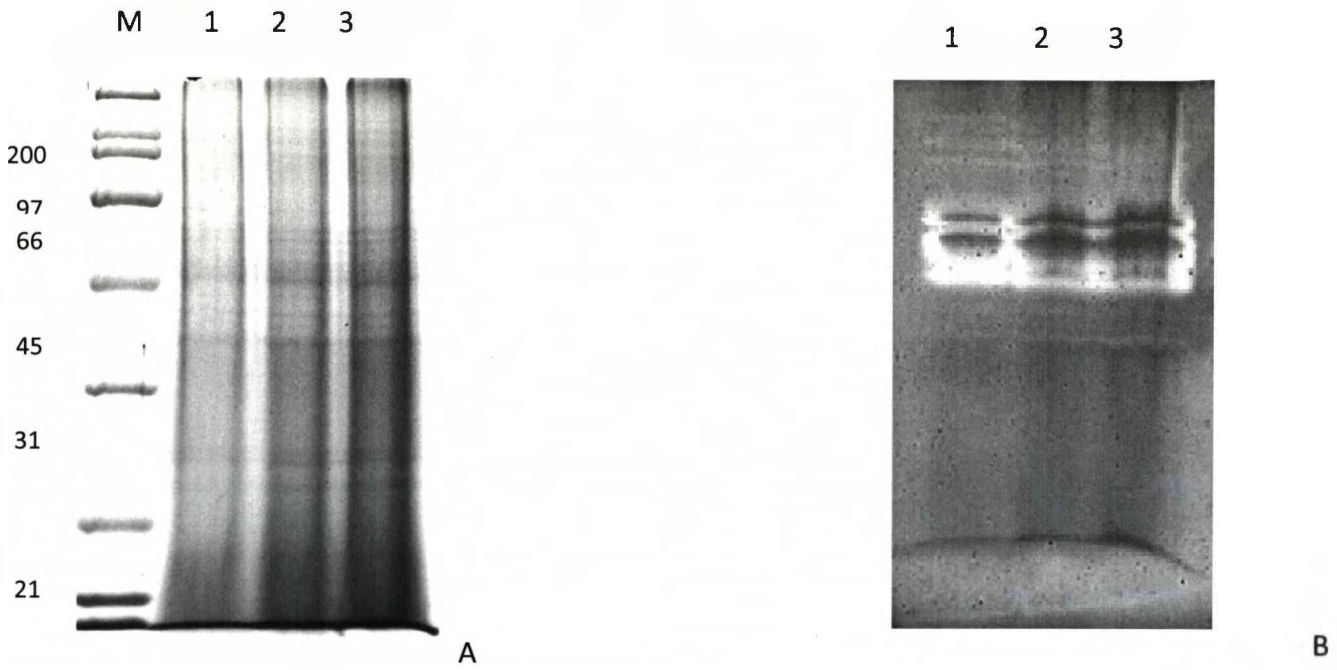


Figure 4.3 Protein Extract from cellulose bait using UTUCHAPS/CTAB buffer

An example of proteins extracted from cellulose bait by gentle agitation in UTUCHAPs/CTAB buffer for 1 h.

A. Lanes 1-3 loaded with 5, 10 & 15 µL of protein extract (protein extracts were not quantified due to commtaminating substances) and separated by SDS-PAGE. The gel was stained with Coomassie and destained with 10 % (v/v) acetic acid; 10 % (v/v) methanol. The molecular weight marker was Bio Rad broad range protein marker.

B. Lane 1-3 protein loaded with 5, 10 & 15 µL of protein extract and the gel developed as a zymogram. 0.1 % CMC was incorporated into the 10 % resolving portion of a polyacrylamide gel. Following separation the gel was washed 2 x with dH_2O and incubated overnight in 0.1 M sodium phosphate buffer, 1 % Triton X-100, pH 6.4. The gel was stained with 0.1 % Congo red (Sigma) and destained with 1 M NaCl. A protein marker was not used for this gel

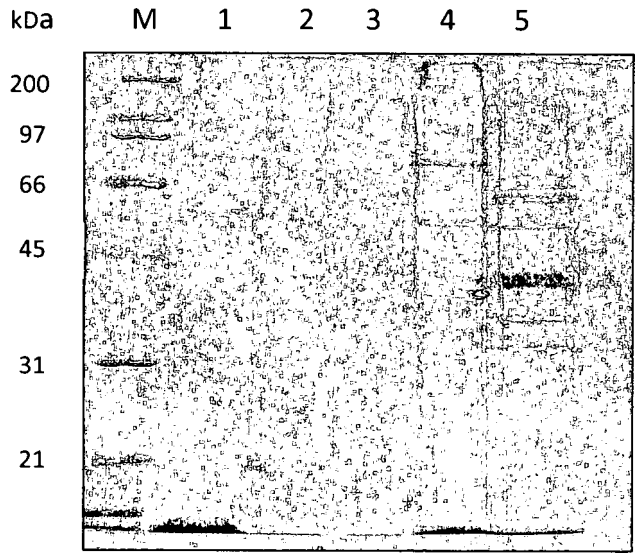


Figure 4.4 One-dimensional SDS-PAGE of proteins extracted using different buffers from colonised cellulose baits.

10 µL of each protein extract was separated by SDS-PAGE. Gels were stained with Coomassie and destained with 10 % (v/v) acetic acid; 10 % (v/v) methanol.

Lane 1, Protein extracted from cotton cellulose bait by shaking with urea, thiourea and CHAPS (UTUCHAPS) buffer for one hour and protein precipitation with trichloroacetic acid (TCA) (20%).

Lane 2. Protein extracted from cotton cellulose bait by shaking in 1 M NaCl and protein precipitation with TCA.

Lane 3. Protein extracted from cotton cellulose bait by shaking without buffer precipitation with trichloroacetic acid.

Lane 4. Protein extracted from cotton cellulose bait by shaking in UTUCHAPS buffer and precipitation with trichloroacetic acid in acetone.

Lane 5. Protein extracted from cotton cellulose bait by shaking in 1 M NaCl and precipitation with TCA in acetone.

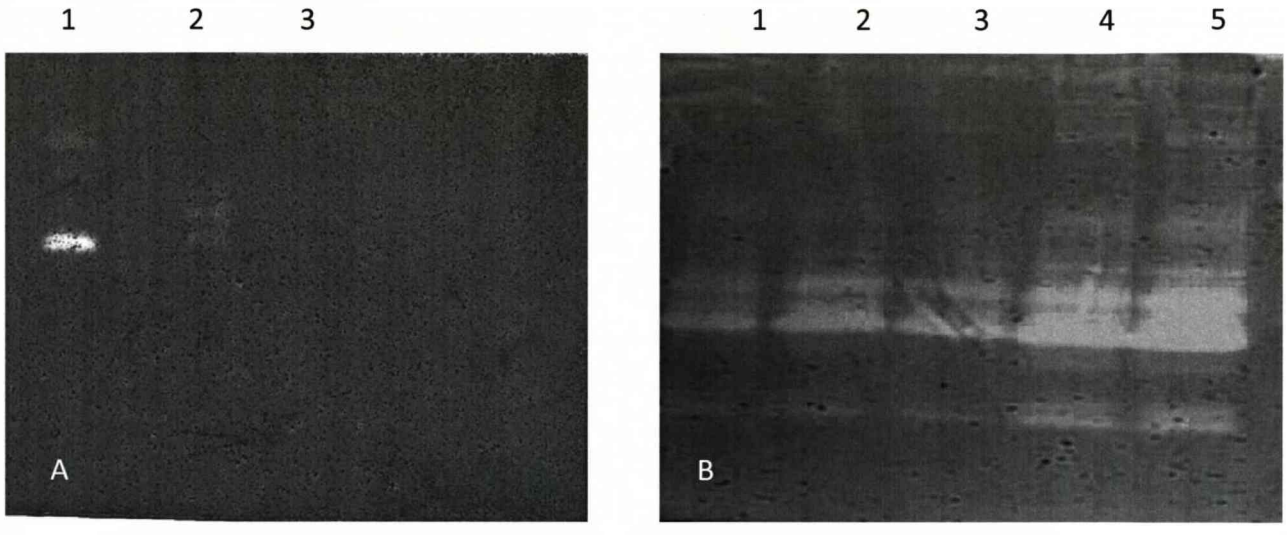


Figure 4.5 Zymogram analysis of protein extracted from a cellulose bait

A, Protein extracted from cellulose baits using NaCl **(1)**, UTUCHAPS **(2)** and UTUCHAPS/CTAB **(3)**. All extracts precipitated with TCA in acetone and resuspended in PBS 10 µL of each extract separated by SDS-PAGE.
B, Protein extracted with NaCl, 0.5 M **(1)**; 1 M **(2)**; 2 M **(3)**; 3 M **(4)**; 4 M **(5)**. All extracts precipitated with TCA in acetone and pellets resuspended in PBS. 25 µg of protein loaded in each lane.
Both A and B were developed as zymograms. Following electrophoresis, gels were washed in dH_2O and incubated overnight in 0.1 M sodium phosphate, 1 % Triton X-100, pH 6.4. Gels were stained with 0.1 % Congo red and bands of activity visualised by destaining with 1 M NaCl.
Protein standards were not included therefore the molecular weight of active proteins could not be estimated.

Development of protein extraction methods was based on routine use of 1D SDS-PAGE and zymogram analysis. Ultimately it was anticipated that proteins would be separated by 2D PAGE, as has previously proved successful with metaproteomic analysis elsewhere (Wilmes & Bond, 2004). Figure 4.6 shows proteins extracted from cotton string baits using UTUCHAPS/CTAB buffer and separated by 2D PAGE. The streaking and smearing observed on 1D gels, was amplified by 2D PAGE (Figure 4.6). A number of factors could be responsible for the poor quality and resolution. Firstly, co-extraction of DNA in the sample was addressed by including an endonuclease in the sample preparation procedure. Standard DNAase could not be used as it is inactivated by the urea that was present in the buffer, consequently Benzonase®, which is not affected by urea was used (as per manufacturers' instructions). However, the resolution of the protein bands was not improved by this modification to the protocol. A more likely explanation was the co-extraction of lipids, polysaccharides and or pigments. In plant proteomics, a phenol extraction step is incorporated into the sample preparation to overcome these problems (Faurobert, 2007), and this was included here (Figure 4.6). Although this significantly decreased streaking and background staining, it also reduced the number of stained protein spots visible on the gel. It is unclear whether this was due to protein removal by phenol extraction or sample-sample variations. In addition, after phenol extraction, bands of activity on zymograms were no longer visible and we were unable to determine whether this was due to loss of protein through the phenol extraction or loss of activity. Phenol is commonly used in DNA extractions for the removal of protein therefore protein could be lost at significant levels at this point. Further difficulties included the quantification of protein, following use of a number of quantification methods, the most reliable was found to be the 2D Quant kit, GE Healthcare, however even using this method, when sufficient amounts of protein were loaded onto IPG strips, characteristically this did not correspond with the number of spots visualised on the resolving gel and the number of bands visualised with 1D SDS-PAGE is not reciprocated with 2D analysis. It is possible that total protein quantification was accurate but a complex mixture of proteins in low abundance may be visible on 1D SDS-PAGE where multiple proteins

with the same molecular weight run to the same point on a gel but may not be visible at the high quantities required by individual spots. Ultimately it proved too difficult to improve definition and separation using 2D PAGE and the alternative method of chromatography was sought.

Following the difficulties with protein separation using gel based methods discussed above and personal communication with Dr Mark Wilkinson, University of Liverpool chromatography separation was employed. Proteins were extracted from cellulose baits as described (4.2.1; Figure 4.1) and dialysed against 50 mM Tris-HCl, pH 8 to remove NaCl. The extracted proteins were separated by anion exchange chromatography (AnIEX). All fractions were subjected to zymogram analysis (data not shown), and fractions showing bands of endoglucanase activity were pooled. 3 M NaCl was added to the pooled anion exchange fractions applied to a phenyl sepharose column equilibrated with 3 M NaCl and eluted in sodium phosphate buffer for fractionation by HIC. Following zymogram analysis of HIC separated proteins active fractions (data not shown) were subsequently pooled and concentrated to 20 µL using a 5 kDa cut off centrifuge concentration tube (Pierce) and dialysed in ammonium bicarbonate (10 mM) overnight. This resulted in a final volume of 100 µL, of which 50 µL was removed and dried under centrifugal evaporation. Laemmli buffer was added (1X) and proteins separated by 1D SDS-PAGE. It can be seen from Figure 4.7 that the use of chromatography produces a 'cleaner' protein extract, than using precipitation methods. Several bands can be seen clearly with an extremely reduced amount of background staining and smearing on the gel.

Although chromatographic separation was initially viewed as a means of protein separation, it proved to act as a method of concentration, while 'cleaning up' the protein extract. However, protein was lost at each step and polyacrylamide gel electrophoresis results in losses of protein and peptides following tryptic digestion, consequently a 'shotgun proteomics' approach was adopted. As anion exchange chromatography proved effective at enriching protein in the sample, it was used as a concentration step following extraction.

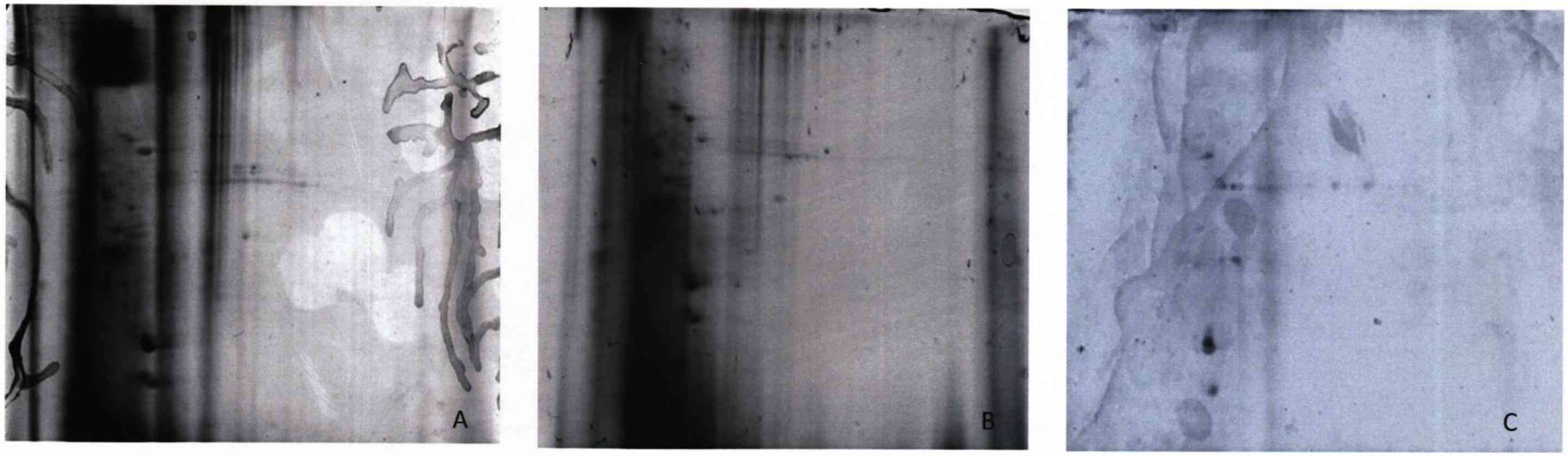


Figure 4.6 Two dimension polyacrylamide gel electrophoresis (2D PAGE) of proteins extracted from cellulose bait

2D SDS-PAGE of proteins extracted using gentle agitation in UTUCHAPS/CTAB buffer for 1 h at 4 °C **(A)**, Protein extract of A with the addition of Benzonase® **(B)** and Protein extract from a different date with the addition of a phenol extraction step **(C)**. A and B stained with colloidal Coomassie (Severn biotech) as per manufacturer's instructions and C stained using a double silver stain method (Heukshoven & Dernick, 1985).

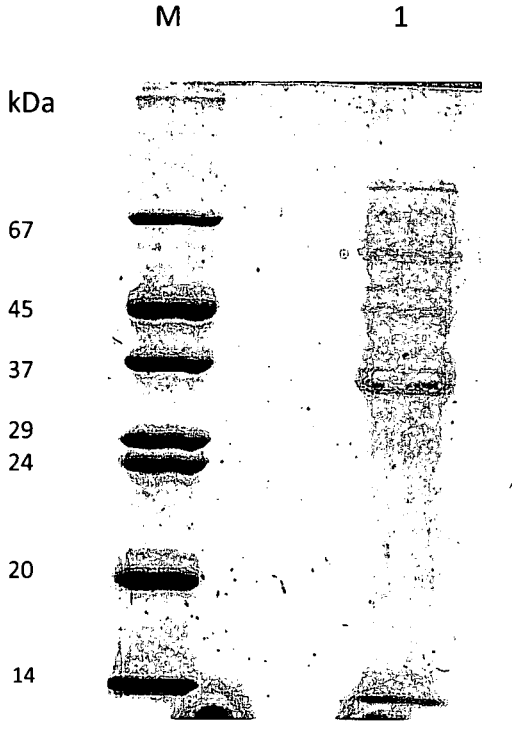


Figure 4.7 1D SDS-PAGE analysis of AnIEX and HIC separted protein fractions

Protein fraction separated by anion exchange chromatography were analysed for CMCase active proteins as measured by zymograms. Active fractions were pooled and further separated by hydrophobic interaction chromatography. When analysed further using zymograms active fractions were pooled and separated by SDS-PAGE **(1)**. The SDS-7 molecular weight protein marker (Sigma) was used.

4.3.4 Final Protein extraction and fractionation method

Method development and optimisation for extraction, separation and analysis of biofilm proteins led to the following final method:

Half the weight of 67.7 g (wet weight) of cellulose bait (returned October 2007) was added in volume (34 ml) of 4 M NaCl and agitated gently for 1 h at 4°C. All liquid was removed as described (figure 4.1) and centrifuged at 5000 x g for 30 min at 4°C. The supernatant was removed and dialysed at 4°C in 2 M NaCl (2 L) overnight. The dialysis membrane was rinsed with ddH_2O and placed in 2 L Tris-HCl (50 mM), pH 8.2 for 8 h and then for a further 8 h and 4 h in fresh buffer. The resultant 90 mL dialysed protein extract was applied to an anion exchange column, washed with Tris-HCl (50 mM), pH 8.2 and all protein eluted in a single fraction (1.5 mL) with Tris-HCl (50 mM), pH 8.2 containing 1 M NaCl. 10 µL of the concentrated protein extract was analysed by 1D SDS-PAGE (method 4.2.3.2). Figure 4.8 shows total protein extraction from cellulose bait concentrated by AnIEX and separated by 1D SDS-PAGE. A large number of distinct clear bands can be seen and low yields of protein found with previous methods were alleviated using this 'in one' clean up/ concentration method, also enabling an increased amount of starting material to be used.

The protein extract was quantified using the Sigma BCA method and 500 µg of this was digested with trypsin as described (4.2.5.1). Peptides were desalted using a Sephadex G25 column and peptides were then separated by cation exchange chromatography (CatIEX). In total, twelve fractions were collected. Following analysis of the elution profile, one fraction (fraction 7) was chosen and concentrated by centrifugal evaporation, and peptides further separated by reverse phase chormatography (RPC). A total of 7 fractions were collected (figure 4.9). Each fraction was applied to an UltiMate nano-liquid chromatography column (LC Packings) connected to a Waters (Manchester, UK) Q-TOF Micro electrospray tandem mass spectrometer and partial peptide sequences were determined by manual interpretation of MS/MS data using the PepSeq software within the MassLynx package. Peptides were provided for six of the fractions (2-7) and a total of 39 peptide sequences were deduced. Unfortunately due to time constraints only one CatIEX fraction was examined further.

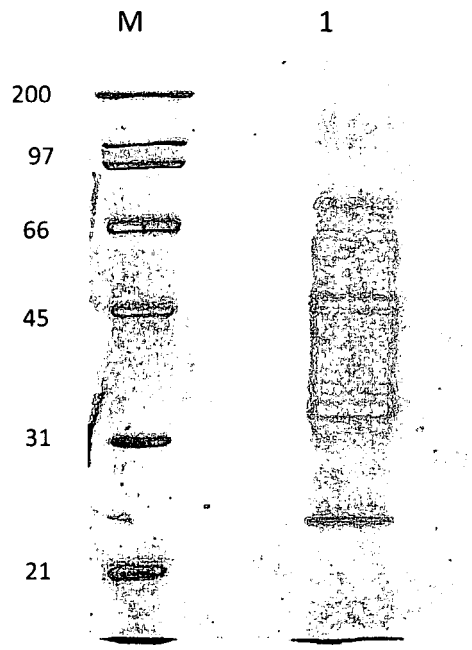


Figure 4.8 1D SDS-PAGE of protein extracted from colonised cellulose using NaCl and concentrated using anion exchange chromatography.

10 µL concentrated protein preparation, eluted from an AnIEX column was separated by SDS-PAGE using a 10 % resolving polyacrylamide gel and a 4 % stacking polyacrylamide gel. The gel was stained with Coomassie and destained with 10 % methanol; 10 % acetic acid. A broad range molecular weight marker (BioRad) was used.

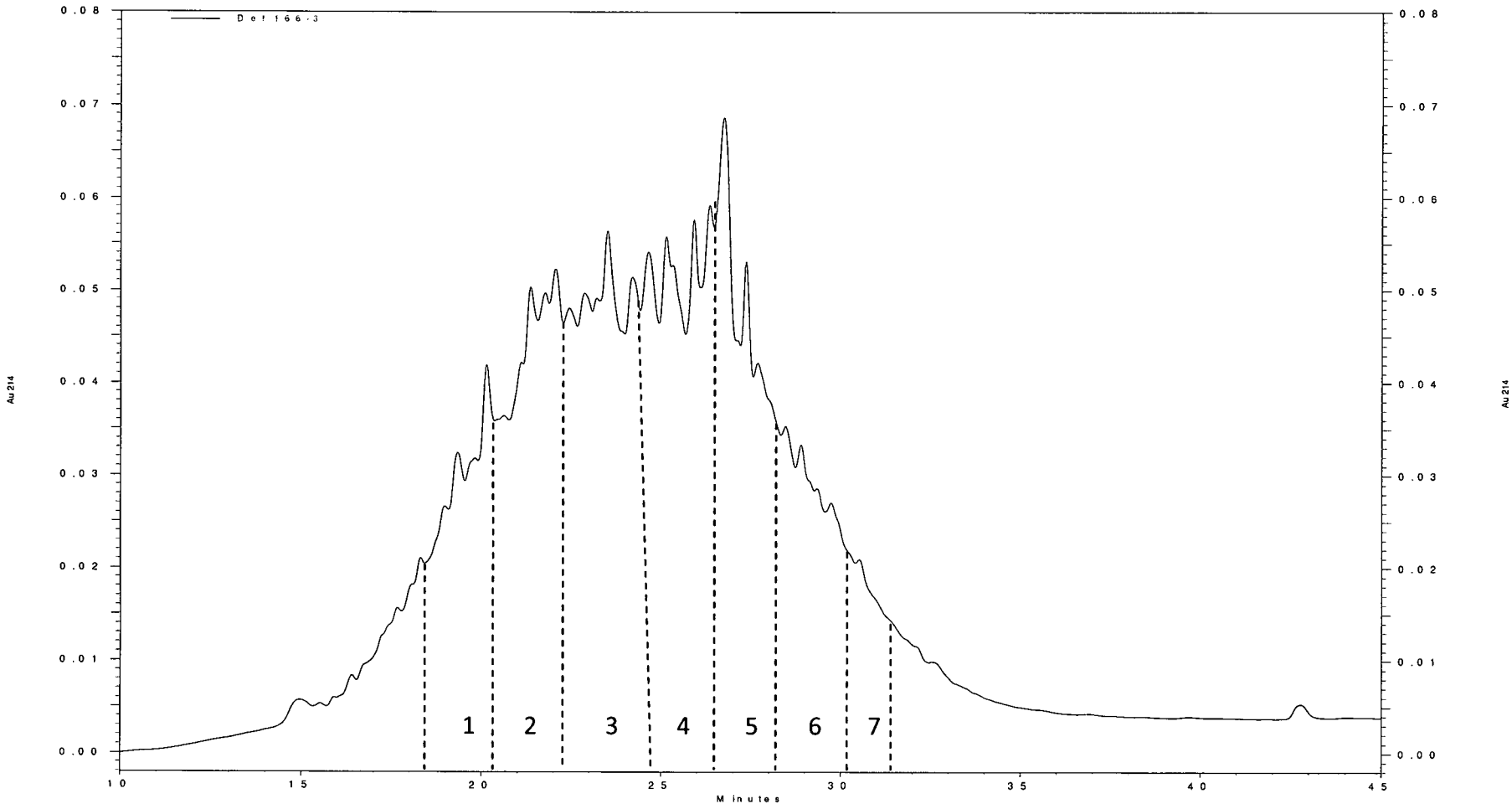


Figure 4.9 Reverse phase chromatography Elution profile

Elution profile for CatIEX fraction 7 (indicated 1-7) separated by reverse phase chromatography (RPC). Peptides were applied to a PepMap C18 Reverse Phase-HPLC column (100 x 2.1 mm) (Applied Biosystems) which had been equilibrated with 0.1 % TFA. Peptides were separated with a 30 min gradient of 0-64 % acetonitrile in 0.1 % TFA. Elution was monitored at 214 nm.

4.3.5 Protein bioinformatics

One of the main issues in the field of proteomics, and especially metaproteomics, is the paucity of sequence data representing the uncultured portion of microorganisms in the reference databases. Therefore the 454 pyrosequencing data retrieved from the same sample site (Chapter3) was used as a database to search peptide sequences using a tblastn search. All matching 454 contigs were then individually searched using a blastx against the NCBI nr database to deduce a function for the proteins matched to the peptides, based on sequence similarity to the known database (Table 4.1). Additionally peptide sequences were searched using blastp against a downloaded version of the non-redundant NCBI protein database (Table 4.2).

When the peptides were searched against 454 pyrosequencing data from the same site (Table 4.1), 38 of the 39 peptide sequences provided a match. However these matches are not entirely convincing; only two peptides had 100 % matches to contigs in the dataset (peptide 4 & 8). When these contigs were used as a blastx query against the NCBI nr-protein database, contigs 01723 showed greatest homology to an outer membrane adhesion-like protein from *Shewanella woodyi* while contig 17978 (peptide 8) did not show homology to any sequence in the nr-database. The closest match to a glycosyl hydrolase protein was provided by peptide 7 which had 5/7 amino acids matched to contig 25648, and when compared to the nr-database the contig showed most homology to a xylanase produced by *Cytophaga hutchonsii*. No larger peptides showed close similarity to contigs in the Irish Sea 454 dataset.

Table 4.1 Irish Sea cellulose bait derived peptides compared to the Irish Sea 454 pyrosequencing dataset

RPC fraction	Peptide number	Peptide sequence	Contig matched	Number AA's matched (% identitiy)	Top hit of contigs (BLASTX against nr database)
2	1	DTDGDGVR	06693	7/8 (88 %)	No significant hit
	2	TSLTVEDAQAR	00052	6/9 (67 %)	No significsnt hit
	3	GDDFTADDVAR	08684	7/11 (64 %)	No significant hit
3	4	QYVDQGGSGLVR	01723	6/6 (100 %)	Outer membrane adhesion-like protein (*Shewanella woodyi*)
	5	LPGGAGGGWDQTR	09029	7/11 (64 %)	exonuclease ABC, A subunit (*Pseudoalteromonas atlantica* T6c)
	6	PTDPNSDFLR	06894	6/9 (67 %)	HflC protein (Teredinibacter turnerae T7901)
	7	FEGSFDAFK	25648	5/7 (71 %)	Xylanase (Cytophaga hutchinsonii ATCC 33406)
	8	SNGNDRVLK	17978	6/6 (100 %)	No significant hits
	9	ELGDELCQR	16051	6/9 (67 %)	ABC1 family protein (Roseobacter sp. CCS2)
	10	SVDEYGK	12073	5/7 (71 %)	Colleganase (Vibrio campbellii AND4)
4	11	NVLVLAEEK	23113	4/9 (44 %)	acetyl-CoA carboxylase, carboxyl transferase, alpha subunit (Teredinibacter turnerae T7901)
	12	LPGLGGDVMAK	12465	6/9 (67 %)	ferrous iron transport protein B [unidentified eubacterium SCB49]
	13	FDTEPDR	06181	5/6 (83 %)	hypothetical protein ISM_04515 [Roseovarius nubinhibens ISM]
	14	EMLEEWNTR	12617	4/9 (44 %)	oligoribonuclease [Saccharophagus degradans 2-40]
	15	AKWFKVAK	15022	4/6 (67 %)	ferredoxin reductase [Myxococcus xanthus DK 1622]
	16	ADVVNPPR	05059	5/8 (63 %)	2,3-dihydroxybenzoate-AMP ligase [Haliangium ochraceum DSM 14365]
	17	VDNAVYDAFK	12853	6/9 (67 %)	secreted protein [Pseudoalteromonas haloplanktis TAC125]
	18	LEEW	No match	N/A	N/A
	19	PDGPLFGR	23927	5/6 (83 %)	Putative site-specific recombinase [Burkholderia glumae BGR1]
	20	YAEEGLDELR	15417	5/8 (83 %)	cytochrome P450, putative [Talaromyces stipitatus ATCC 10500]
	21	MPGLGMQA	02073	4/6 (67 %)	GTP diphosphokinase (guanosine 3',5'-polyphosphate synthase) [Cytophaga hutchinsonii ATCC 33406]
	22	SPTDPNADFLR	24695	8/11 (73 %)	conserved hypothetical protein [Rhodobacterales bacterium Y4I)

	23	NPSVDLGDLR	25882	6/10 (60 %)	conserved hypothetical protein [uncultured beta proteobacterium CBNPD1 BAC clone 578]
	24	VNAMQGEFPVR	10690	5/8 (63 %)	predicted protein [Uncinocarpus reesii 1704]
5	25	LPGTELYGK	19975	5/8 (63 %)	PAS domain S-box protein [Sphaerobacter thermophilus DSM 20745]
	26	EAFGSFDAFK	05260	6/9 (67 %)	serine/threonine kinase [Cellvibrio japonicus Ueda107]
	27	AEDDGLFFR	12072	6/8 (75 %)	No significant matches
	28	TAEDNGLFFR	11372	7/10 (70 %)	5,10-methylenetetrahydrofolate reductase [Flavobacteriales bacterium ALC-1]
6	29	FDPALVDLYR	12480	6/9 (67 %)	aspartyl-tRNA synthetase [Pseudoalteromonas atlantica T6c]
7	30	TLDNDLMLLK	04743	6/8 (75 %)	hypothetical protein bcere0004_27630 [Bacillus cereus BGSC 6E1]
	31	PLPEVLR	07751	6/7 (86 %)	folylpolyglutamate synthase [Microscilla marina ATCC 23134]
	32	TLDNDLFLLK	12178	6/9 (67 %)	glycoside hydrolase family protein [Solibacter usitatus Ellin6076)
	33	SFPLPEVLR	07969	6/7 (86 %)	LD15203p [Drosophila melanogaster)
	34	FADGPDGWPMR	17422	5/10 (50 %)	Hypothetical protein NGR_b08590 [Rhizobium sp.)
	35	PGFVVTEFR	25306	4/8 (50 %)	Peptidase S10 serine carboxypeptidase [Beijerinckia indica subsp. indica ATCC 9039]
	36	FELPALPYER	02846	7/8 (88 %)	superoxide dismutase [Flavobacterium psychrophilum JIP02
	37	FPDNLDDLPSR	03111	7/10 (70 %)	No significant hits
	38	LTEVAEAEGWR	26515	5/10 (50 %)	No significant hits
	39	TTFGLPDLR	19307	8/9 (90 %)	phage tail Collar [Silicibacter sp. TM1040)

Matches and assignments for Peptides searched against the Liverpool Bay 454 sequence database using a tblastn search. All contigs with matches were searched against the nr database using a blastx search to identify homology of translated 454 sequences to known proteins

When peptides were compared to a downloaded version of the NCBI nr database using a BLASTP search, 38 of the 39 peptides gave matches to reference proteins, of which eight were against ABC substrate binding transporter proteins. Some of these are particularly large peptides, peptide 3 (11/11-100 %) and peptide 22 (11/11-100 %) being the longest peptides with the most homology to reference proteins. There are also two peptides (1 & 33) matching at high levels of similarity (7/8 aa & 8/9 aa respectively) to lipoproteins which have previously been described to play a role in cellulosome-like protein complexes in Gram negative bacteria (Weiner *et al.*, 2008; Yang *et al.*, 2009). A PBPb superfamily protein described as having a substrate binding role and interacting with a membrane bound complex (peptide 5) and a PBPb extracellular solute binding protein (peptide 9) are also represented showing a high level of similarity (10/12 (83 %) and 8/9 (89 %) respectively). In addition, two peptides (11 & 23) showed high levels of similarity (83 % & 100% respectively) to an amino peptidase and an oligopeptidase. Notably, two phage proteins were tentatively identified; peptide 16 (7/7) and peptide 39 (9/9) showed high level matches (100 %) to a *Prochlorococcus* (A marine Cyanobacteria) phage T4-like base plate initiator and *Clostridium cellulolyticum* (known cellulosome producing species) phage tail collar respectively. It would be expected that in a marine biofilm community, bacteriophage would play significant roles in bacterial population dynamics.

One of the greatest obstacles to identifying peptides from complex mixtures in liquid based metaproteomics is to match more than one peptide to a protein. This is easier with gel based methods, as it is known that the peptides have originated from the same protein. It is clear however that a large proportion of peptides have produced high level similarity matches to extracellular and membrane complex proteins, which was the target for the extraction method. Although no clear matches to glycosyl hydrolases were detected, proteins involved in attachment to substrates and transport of the products of degradation of polysaccharides are present (i.e. xylose transporter protein). Given that the most abundant peptides are detected with the MS platform used here, it can be suggested that ABC transporter proteins play a significant role in competitive biofilm communities. Confidence can also be drawn from the

matches provided with a large proportion having high level similarity to known marine bacteria species. For example, *Vibrio* spp., *Roseobacter* spp. and three high level similarity matches to proteins from *Phaeobacter gallaeciensis* (Table 4.2).

Table 4.2 Irish Sea cellulose bait derived peptides compared to the NCBI-NR database.

RPC fraction	Peptide number	Peptide sequence	Closest match	Function	Number peptides matched (% identity)
2	1	DTDGDGVR	Chlamydia pneumonia (NP_224795.1)	Oligopeptide binding lipoprotein	7/8 (87 %)
	2	TSLTVEDAQAR	Nocardioides sp (YP_925195.1)	Hypothetical Conserved protein (protein kinase)	9/11 (81 %)
	3	GDDFTADDVAR	Roseobacter sp (ZP_01753347.1)	Oligopeptide ABC transporter, periplasmic substrate binding	11/11 (100 %)
3	4	QYVDQGGSGLVR	Ruegeria pomperoyi (YP_167174.1)	Branched-chain aa ABC transporter, periplasmic substrate binding	10/11 (91 %)
	5	LPGGAGGGWDQTR	Vibrio vulnificus cmc06 (NP_761553.1)	PBPb superfamily-bind substrate and interact with membrane bound complex	10/12 (83 %)
	6	PTDPNSDFLR	Roseobacter sp CCS2 (ZP_01751196.1)	Xylose ABC transporter protein	10/10 (100 %)
	7	FEGSFDAFK	Flavobacteriales bacterium HTCC2170 (YP_001850843.1)	Hypothetical conserved PGRP-peptidoglycan (PGN)reconition, sometimes hydrolyse PGN of bacterial cell walls	8/8 (100 %)
	8	SNGNDRVLK	Mycobacterium marinum (YP_001850843.1)	Conserved hypothetical protein	8/9 (89 %)
	9	ELGDELCQR	Phaeobacter gallaeciensis BS107 (ZP_02147056.1)	PBPb extracellular solute binding	8/9 (89 %)
	10	SVDEYGK	Platynereis dumerilli (marine annelid) (CAJ38792.1)	Notch protein	7/7 (100 %)
4	11	NVLVLAEEK	Ruminococcus gnavus ATCC 29149 (ZP_02040505.1)	Hypothetical conserved aminopeptidase domain	7/9 (78 %)
	12	LPGLGGDVMAK	Trichodesmium erythraeum (Cyanobacteria) (YP_723883.1)	TRAP dicarboxylate transporter protein	9/11 (82 %)
	13	FDTEPDR	Sinorhizobium meliloti (NP_386436.1)	Cold shock transcription regulator protein	7/7 (100 %)
	14	EMLEEWNTR	Alcanivorax borkumensis SK2	Formamidase	9/9 (100 %)

			16(YP_693698.1)		
	15	AKWFKVAK	Bacillus weihenstephanensis KBAB4 (Y18P_001642580.1)	Multidrug (MATE) efflux pump	7/8 (88 %)
	16	ADVVNPPR	Prochlorococcus phage (YP_214343.120)	T4-like baseplate wedge initiator	7/7 (100 %)
	17	VDNAVYDAFK	Burkholderia ambifaria (YP_001807786.1)	Acriflavin resistance protein	8/9 (89 %)
	18	LEEW	No match	N/A	N/A
	19	PDGPLFGR	Magnetospirillum gryphiswaldense MSR-1ABN79615.1	Hypothetical protein	7/7 (100 %)
	20	YAEEGLDELR	Rhodobacterales bacterium HTCC 2654 (ZP_01014460.1)	Branched-chain aa ABC transporter protein	10/10 (100 %)
	21	MPGLGMQA	Bos Taurus (cattle) (XP_874420.2)	Regulator of G-protein signalling 22	7/7 (100 %)
	22	SPTDPNADFLR	Phaeobacter gallaeciensis BS107 (ZP_02146583.1)	D-xylose ABC transporter, substrate binding	11/11 (100 %)
	23	NPSVDLGDLR	Methylocella silvestris BL2 (ZP_02947450.1)	Oligoendopeptidase, pepF/M3 family	9/9 (100 %)
	24	VNAMQGEFPVR	Oryza sativa Japonica group (BAC84664.1)	Unknown protein	8/11 (73 %)
5	25	LPGTELYGK	Citrobacter koseri ATCC BAA-895 (YP_001454448.1)	Flagella biosynthesis protein FlhA	7/9 (78 %)
	26	EAFGSFDAFK	Exiguobacterium sibiricum 255-15 (YP_001813355.1)	Superoxide dismutase	10/10 (100 %)
	27	AEDDGLFFR	Hoeflea phototrophica DFL-43 (ZP_02165233.1)	ABC aa transporter, periplasmic substrate-binding protein	9/9 (100 %)
	28	TAEDNGLFFR	Sagittula stellata E-37 (ZP_01747041.1)	ABC branched aa transporter, periplasmic substrate binding protein	10/10 (100 %)
6	29	FDPALVDLYR	Thermobifida fusca (soil cellulolytic bacteria) (YP_289279.1)	Actinorhodin polyketide β-ketoacyl synthase α-subunit	7/10 (70 %)
7	30	TLDNDLMLLK	Human (gi|162330095)	Chain A, Human Mesotrypsin Complexed With Bovine Pancreatic Trypsin Inhibitor(Bpti).	8/10 (80 %)

31	PLPEVLR	Parabacteroides merdae ATCC 43184 (ZP_02033198.1)	Hypothetical membrane protein	7/7 (100 %)
32	TLDNDLFLLK	Culex quinquefasciatus (XP_001868151.1)	Trypsin 7	8/10 (80 %)
33	SFPLPEVLR	Anaeromyxobacter dehalogenans 2CP-1 (ZP_02321770.1)	Membrane lipoprotein	8/9 (90 %)
34	FADGPDGWPMR	A-proteobacteria HTCC 225 (ZP_01448759.1)	OmpA protein	7/7 (100 %)
35	PGFVVTEFR	Schizosaccharomyces pombe 972h (NP_592771.1)	Predicted short chain dehydrogenase	7/9 (78 %)
36	FELPALPYER	Marinobacter sp ELB 17 (ZP_01736020.1)	Superoxide dismutase, Fe	9/10 (90 %)
37	FPDNLDDLPSR	Phaeobacter gallaeciensis BS107 (ZP_02146704.1)	AFG1-like ATPase	8/11 (73 %)
38	LTEVAEAEGWR	Oceanicola granulosus HTCC2516 (ZP_01157244.1)	Periplasmic ABC substrate binding transporter protein, signal peptide	9/11 (82 %)
39	TTFGLPDLR	Clostridium cellulolyticum (ZP_01574157.1)	Phage tail collar	9/9 (100 %)

Matches and assignments of peptides searched against the NCBI nr database using Blastp

4.3.6 Metaproteomic analysis of protein extracted from chitin baits

The methods developed for extraction, separation and analysis of proteins from biofilm communities of cellulose baits, described above, were adapted and applied to chitin biofilm communities.

Proteins were extracted according to the final optimised method (section 4.3.4) with one modification: chitin does not significantly retain water so the NaCl concentration was reduced to 2 M. A total of 15.8 g chitin which had been tethered to the Liverpool Bay bouy B site for one month (August-September 2008) was added to an equal volume of 2 M NaCl and shaken at 4°C, 1 h. All liquid was extracted by centrifugation of screw cap tubes at 5000 x g, 20 min and removal of supernatant which was further centrifuged at 20000 x g for 30 min.

It was apparent that the complexity of the protein mixture extracted from chitin was less than that of the cellulose biofilm. Therefore, proteins were separated by AnIEX, but instead of collecting all proteins in a single fraction, a gradient of 0-1 M NaCl was used and three 0.5 mL fractions were collected. Laemmli buffer was added to 10 µL of each fraction and proteins separated by 1D SDS-PAGE (Figure 4.10). Proteins were also run on a zymogram, with ethylene glycol chitin as the substrate. It was clear from figure 4.10 that all activity was in one fraction (fraction 1) which was subsequenctly concentrated using a centrifuge concentrator MWCO 5kDa (Vivaspin Sartorius), proteins separated by 1D SDS-PAGE, and bands excised from the gel and digested with trypsin. Peptides were subject to LC/MS/MS as previously described, however no peptide sequences could be deduced from the mass spectra.

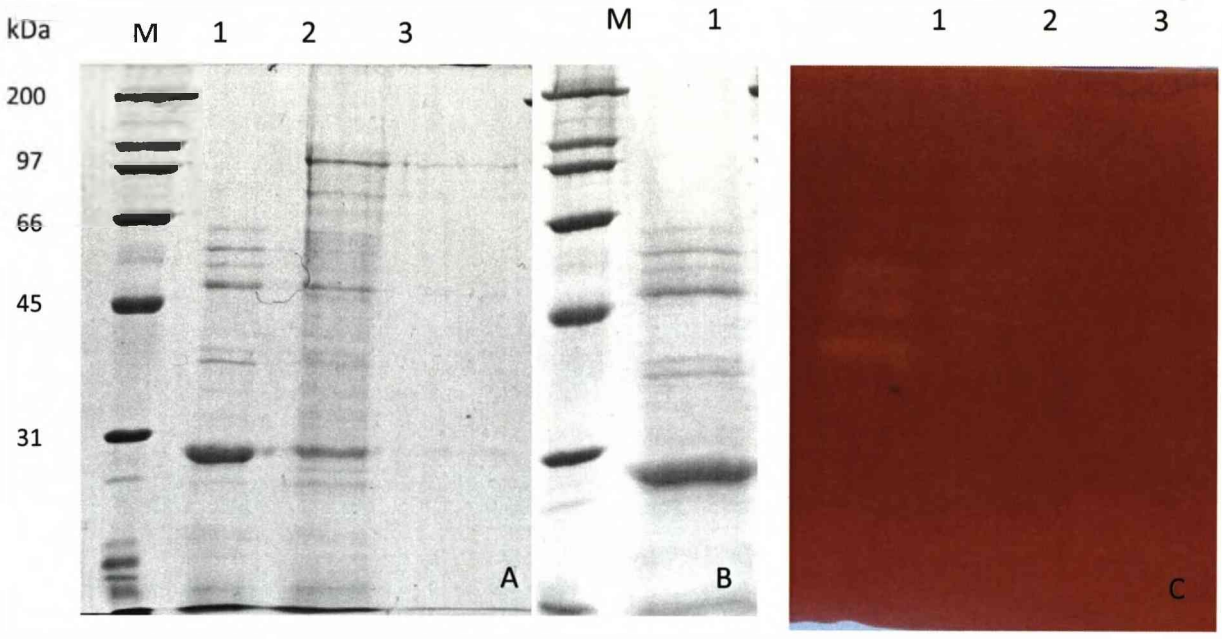


Figure 4.10 SDS gel electrophoresis and of protein extracted from chitin bait

Proteins were extracted from chitin and concentrated by AnIEX. Three fractions were collected and separated by 1D SDS-PAGE **(A)**. Lane 1, 2 & 3 represent fractions 1, 2 & 3 respectively. Proteins were separated by 1D SDS-PAGE and developed as a zymogram **(C)**. Fraction 1 was found to contain chitinase activity and was subsequently concentrated using a 5 kDa cut off concentration tube and the sample separated by 1D SDS-PAGE **(B)**. A broad range molecular weight marker (BioRad) was used for **A** and **B**. A molecular weight marker was not used for **C** therefore molecular weight of active proteins could not be estimated.

4.4 Discussion

Extraction of proteins from environmental samples is clearly the pivotal step in any metaproteomics project as it dictates which proteins will form the basis of the investigation. However, downstream processing methods also impact on the success of identification. There are clear advantages for gel based proteomics such as the additional information provided (molecular weight and pI) and the confidence in data on peptides known to have originated from the same protein band. 2D PAGE has previously been applied in metaproteomic studies (Wilmes & Bond, 2004; Kan *et al.*, 2005; Klassens *et al.*, 2006), and although retrieval of peptide information from individual spots may provide an advantage with regards to additional molecular weight and isoelectric point information, they all had limited success with protein identification. In addition, gel based proteomics is very time consuming and technically demanding when the quantities of protein starting material are low. Liquid based proteomics on the other hand can be largely automated, especially if the chromatography columns are connected on line with the MS. The disadvantage of this method however is the complexity of peptides present and matching multiple peptides from a single protein within that mixture, particularly for unknown proteins likely to be predominant in environmental proteomics.

Liquid chromatography was found to effectively separate the complex marine biofilm protein extract to manageable levels for MS analysis. Future proteomics and especially metaproteomic studies are likely to favour shotgun based approaches, relying on the digestion of peptide mixtures followed by separation of peptides and subsequent introduction in a tandem mass spectrometer. Single dimension separation has routinely been used with LC MS/MS in proteomics and multi dimensional separation has previously been applied in community proteomic studies. VerBerkmoes *et al.* (2009), combined CatIEX and RPC online with the MS/MS analysis, whereas in this project CatIEX and RPC were performed on peptide mixtures prior to application to the LC MS/MS.

If genomic sequence data is available, spectral peptide mass fingerprint data can enable *in silico* protein identification (Ram *et al.*, 2005; Klassens *et al.*, 2007).

VerBerkmoes *et al.* (2009) utilised a number of genomic databases to provide automated searches of mass spectra using SEQUEST (Eng *et al.*, 1994) and DTASelect/Contrast (Tabb *et al.*, 2002). Wilmes *et al.* (2008) employed the MASCOT search engine (http://matrixscience.com) to search PMF data against known sludge metagenomic databases, obviating the need for *de novo* peptide sequencing. However, with environmental samples containing uncultured species where the constituent species are unknown, peptide sequence data is of much greater value in the absence of comprehensive metagenomic sequence data (Wilmes & Bond, 2004; Lacerda *et al.*, 2007).

When peptides originating from a single protein spot are being analysed, it is difficult to obtain matching hits of all peptides to the same protein (Wilmes & Bond, 2004); this problem is emphasised when peptides cannot be traced back to the protein of origin, as with shotgun approaches. *De novo* sequencing is performed manually on spectral data and it should be remembered that there can be anomalies between some amino acids, such as L and I and in distinguishing between D and N or between E and Q/K. Additionally, combinations of amino acids can yield the same mass number or even the same elemental composition; for example G & G can equal N ($C_4H_6H_2O_2$) and a G & A can equal that of Q ($C_5H_8N_2O_2$) (Standing, 2003) emphasising the requirement for improved accuracy of available MS platforms. These improvements may be addressed with the Ion trap instruments and the introduction of MS^3 whereby three mass analysers are placed in tandem (Olsen & Mann, 2004). Lo *et al.* (2007), Wilmes *et al.* (2008) and VerBekmoes *et al.* (2009) all enlisted an iontrap mass spectrometer, whilst the facilities available to this project were confined to an ESI-QTOF MS. Ion Trap mass spectrometers have shown increased sensitivity and speed, with increased capture efficiency and storage capacity (Olsen & Mann, 2004). Manual interpretation is also time consuming and not practical for high throughput analysis. Reliable bioinformatics would need to be sourced, and environmental metagenomic sequence data retrieved to obtain a meaningful analysis of the biological variability of environmentally derived proteomes.

Multidimensional approaches rely on two or more independent physical properties of proteins or peptides to achieve a higher level of resolution and a higher loading capacity than can be achieved with a single dimension (Delahunty & Yates, 2005). This is due to the fact that the most dominant 'housekeeping' proteins will be detected and low abundant proteins, which are often of interest, may be missed due to the sensitivity of current mass spectrometers (VerBerkemoes *et al.*, 2004).

Recently, 'Top down' MS, which involves direct analysis of intact proteins without proteolytic digestion, has been successful with single proteins (Siuti & Kelleher, 2007). This would aid the identification of post translational modifications or protein sequence variations between strains in a population and between enzyme complexes. Additionally 'middle down' analysis of larger peptide fragments (>3 KDa) has been proposed (Siuti & Kelleher, 2007) as opposed to routinely used 'bottom up' MS of small peptides which can make identification of proteins daunting (Suiti & Keller, 2007). As the average protein can give rise to 20-30 peptides and the average bacterial species will produce 1000 proteins (VerBerkemoes *et al.*, 2004) this can results in a vast number of peptides to be processed by a mass spectrometer. For metaproteomics on typical diverse microbial communities, top down proteomics is a very attractive prospect indeed.

Although the Liverpool Bay sample was relatively complex in terms of species composition (chapter 3), compared to recent metaproteomic studies (Ram *et al.*, 2004) the *in situ* enrichment aimed to limit the diversity of the microbiota. Followed by extensive sample process simplification, a number of functional proteins could be successfully identified. It was initially considered that a targeted approach using polyacrylamide gel electrophoresis separation and MS/MS peptide sequencing of individual spots and bands would provide a means for cellulase and chitinase activity identification within the biofilm community, however the success of multi dimensional separation of tryptic peptides followed by mass spectrometry would suggest that further work on this project be applied to whole protein extracts from string, as the integrity of protein does not need to be maintained with subsequent trypsin digestion. This may lead

to the elucidation of key aspects of functionality within the biofilm community, not limited to primary cellulase and chitinase activity.

This study focused on identifying proteins involved in complex polysaccharide degradation, of which the presence was identified by zymogram analysis. Therefore instead of isolating the total protein complement by lysis of bacteria, extraction was focused on only the extracellular proteins released into the biofilm matrix. This is difficult to achieve without some cell disruption and release of membrane, or cytoplasmic proteins.

Analysis of the metaproteome of polysaccharide colonising microorganisms was hampered by a multitude of difficulties including quantities of starting material, sample variation and co extraction of contaminating components. Ultimately, however it was the man hours involved in separation, MS sensitivity and the time consuming nature of manual peptide sequencing from spectra that limited progress. It is quite clear that future high throughput metaproteomic approaches will need to include solutions to all these difficulties in order to recognise the potential of this field. Although general trends in metabolic activity of the community can be observed, confident assignment of species to those functions within a community of the complexity observed in the cellulose biofilm retrieved for the Irish Sea is still beyond the reach of current metagenomics and metaproteomics in combination. For less complex communities, comprehensive analysis at the strain level of the proteome can be achieved, e.g., the Acid Mine Drainage system (Ram *et al.*, 2005; Lo *et al.*, 2007) and activated sludge (Wilmes *et al.*, 2008).

4.5 Conclusions

- A method for the extraction of community protein from a biofilm colonising cotton cellulose baits was successfully developed.

- The method was subsequently applied to the extraction of community protein from a biofilm colonising crab shell chitin bait.

- Activity of cellulase and chitinase was determined by means of zymogram analysis on extracted community protein from cellulose and chitin colonising biofilms.

- Purification, concentration and fractionation of community protein was achieved by means of a number of chromatographic techniques

- A number of peptides from the cellulose extracted proteins were successfully matched to proteins from the NCBI-nr database with functions involved in carbohydrate metabolism.

Chapter 5

Isolation of Bacteria from Colonised Cellulose retrieved from Liverpool Bay

5.1 Introduction

5.1.1 Bacterial cultivation

Traditionally, culture based methods have been used for the determination of microbial community dynamics. However, the diversity and abundance of microorganisms in natural environments is such that there is no single medium or method to propagate bacterial cells under laboratory conditions. Very early work identified and replicated the conditions under which particular species thrive in their natural habitats using selective media (Jensen *et al.*, 1996; Wery *et al.*, 2003). For example, nutrient rich marine broth 2216 (Zobell, 1941) and nutrient limiting F/2 trace metal media (Guillard & Ryther, 1962) were employed to mimic natural conditions for heterotrophic and oligotrophic marine bacteria. Recently, developments have been made to further recognise the potential of culturing techniques, such as the dilution-to-extinction culturing method (Button *et al.*, 1993; Connon & Giovannoni, 2002; Rappe *et al.*, 2002) which aims to culture bacteria adapted to oligotrophic conditions (Fuhrman & Hagstrom, 2008). Enrichment is also adopted for culturing marine bacteria in which the known requirements of bacteria are applied to their propagation from mixed communities. These include using a requirement for sodium chloride by halophiles, high temperatures for thermophiles or low temperatures for psychrophiles (Wery *et al.*, 2003). However, with the advent of phylogenetic analysis of bacteria using genetic markers it is evident that these culturing methods have made a trivial impact on the extent of diversity present in natural systems.

5.1.2 Aims and Objectives

The aims of this chapter are to characterise the culturable cellulolytic component of a biofilm community colonising cellulose string bait in the Irish Sea and to visualise the biofilm community on the surface of cellulose bait using Scanning Electron Microscopy (SEM).

5.2 Methods

5.2.1 Media, Bacterial strains, Growth and Maintenance

Marine Broth 2216 (Difco) was used for the isolation and propagation of all marine bacterial strains and these were stored in 8% (v/v) DMSO at -80 ^{o} C. Strains were isolated from cotton string tethered at the Liverpool Bay fixed mooring site B (for methods see section chapter 2.1). Cellulose-colonising organisms were extracted for culture by adding a piece of cotton string, approximately 5 cm in length, to 10mL seawater (collected from site B) which had been filter sterilised by passing through a 0.2µM (pore diameter) filter (PALL Corporation) under vacuum. The cotton string was subjected to vigorous shaking and vortexing to detach the biofilm organisms. A series of dilutions (10^{-3}, 10^{-4}, 10^{-5}) were made and 100 µl of each spread onto the marine broth 2216 (Difco) agar plates and incubated at 25 ^{o}C. Colonies that had grown (~1week) were selected on the basis of morphological appearance, and subcultured onto fresh marine broth agar plates.

5.2.2 Screening for endoglucanase activity

The bacterial isolates were inoculated onto a 0.2 µM (pore diameter)membrane filter (PALLCorporation) which had been placed, under aseptic conditions, onto the surface of marine broth 2216 agar plates supplemented with 0.1% Carboxymethyl cellulose (Sigma) and incubated at 25 ^{o}C for 4 days. After incubation, the membranes were removed and discarded. The agar plates were then flooded with 0.1% Congo red (Sigma) and gently shaken for 30 min. The Congo red was discarded and destained by washing 2 X

15 min with 1M NaCl. Plates were viewed using a Syngenta imaging system with GeneSnap software

5.2.3 Polymerase Chain Reaction (PCR) Amplification of 16S rRNA Gene Sequences

PCR reactions were performed in 50 µL volumes containing: 1µL of a 1 in 10 dilution of a colony of each isolate, 0.2 mM each of primers pA and pH (Edwards *et al.*, 1989), 0.2 mM each dNTP, 1 x Phusion HF buffer (Finnzymes), 3 % DMSO, 1 x BSA, 1 unit Phusion™ High-Fidelity DNA Polymerase (Finnzymes) and ddH_2O. PCR cycling conditions were as follows; 98°C for 30 s, 30 cycles of 98°C for 10 s, 30 s at 55°C, 72°C for 20 s and a final extension of 72°C for 8min

PCR amplification products were excised from 1 % agarose gels using a sterile scalpel blade and purified using the Perfectprep® Gel Cleanup kit (Eppendorf) according to the manufacturer's protocol. A-tailing of blunt- ended PCR products generated by amplification with Phusion™ High-Fidelity DNA Polymerase (Finnzymes) was performed followed by ligation and cloning of 16S rRNA gene products into competent *E.coli* JM109 (Promega) cells according to the manufacturer's protocol following the pGEM® -T Easy cloning vector System I (Promega).

5.2.4 Plasmid extraction and sequencing of 16S rRNA gene fragments

Plasmid DNA was extracted from overnight clone cultures (LB broth and 100 µg ml^{-1} ampicillin) using the QIAprep® Spin Miniprep kit (Qiagen) following the manufacturer's protocol, and the insert sequenced in both directions by Macrogen Inc. (South Korea).

Forward and reverse clone insert DNA sequences were assembled into contigs using PreGap 4 and Gap4 software (Staden, 1996) and base calling was visually checked using the sequencing traces. Assembled contigs were subjected to two chimera check packages, RDP Chimera Check (Cole *et al.*, 2005) and Pintail (Ashelford *et al.*, 2005). Sequences were aligned using Greengenes (DeSantis *et al.*, 2006) with the top three

matches and imported into ARB (beta v. 2003-08-22, Ludwig *et al.*, 2004) where the alignment was manually optimised. The alignment was used for the calculation of a neighbour-joining tree within ARB (http://www.arb-home.de/publications.html), using the Olsen correction method with 1000 bootstrap samplings. A maximum-likelihood tree was obtained using PhyML (http://www.atgc-montpellier.fr/phyml/). Default online parameters were applied with the SH-like aLRT setting to give branch support.

5.2.5 Scanning Electron Microscopy (SEM) of colonised cellulose and chitin samples

Samples of cellulose bait were collected from the mooring site in the Irish Sea (Buoy B-January, 2009) and refrigerated until returned to the laboratory. The samples were gently rinsed with ddH_2O and immersed in excess absolute Ethanol (Sigma) which had been pre cooled at -80°C, and the samples returned to -80°C overnight. The samples were then removed and placed into a universal bottle containing pre-cooled absolute ethanol, and again returned to -80°C until required. Specimens were dried from absolute ethanol in carbon dioxide using a Polaron E3000 critical point dryer, glued to stubs, sputter-coated with 60 % gold-palladium in a Polaron E 5100 coater and viewed in a Philips 501B scanning electron microscope at accelerating voltages of 7.2 and 15 kV (Veltkamp *et al.*, 1994). Final sample preparation and primary microscopic examination of the samples was carried out by Cornelis Veltkamp & Carmel Pinnington at the Department of Earth & Ocean Sciences, University of Liverpool.

5.3 Results

5.3.1 Bacterial strain isolation, and screening for endoglucanase activity

Bacterial isolates were collected from cotton string bait moored at the buoy B site (Figure 2.1) in the Irish Sea for one month and returned in February, 2007. Colonies were picked from dilution spread plates based on morphological difference and subsequently a total of 26 isolates were screened for endoglucanse activity. Of 26 isolates, nine were found to be positive for endoglucanase activity (Figure 5.1). A zone of clearing where the β 1,4-glycosidic bond is broken by endo-acting hydrolases resulting in the inability of Congo red to bind to degraded cellulose is clearly visible.

Gram staining was performed on the endoglucanase positive isolates and all were found to be Gram negative rod shaped bacteria. Isolates 54, 56, 58 and 63 grew much slower than the other isolates, forming smaller colonies that were a peach/pink colour. Isolates 47, 48 and 62 formed larger colonies with an opaque colouration and isolate 53, the quickest growing of the nine isolates formed large colonies that produce a deep yellow pigmentation. Isolate 40 formed large colonies with a green/blue pigmentation.

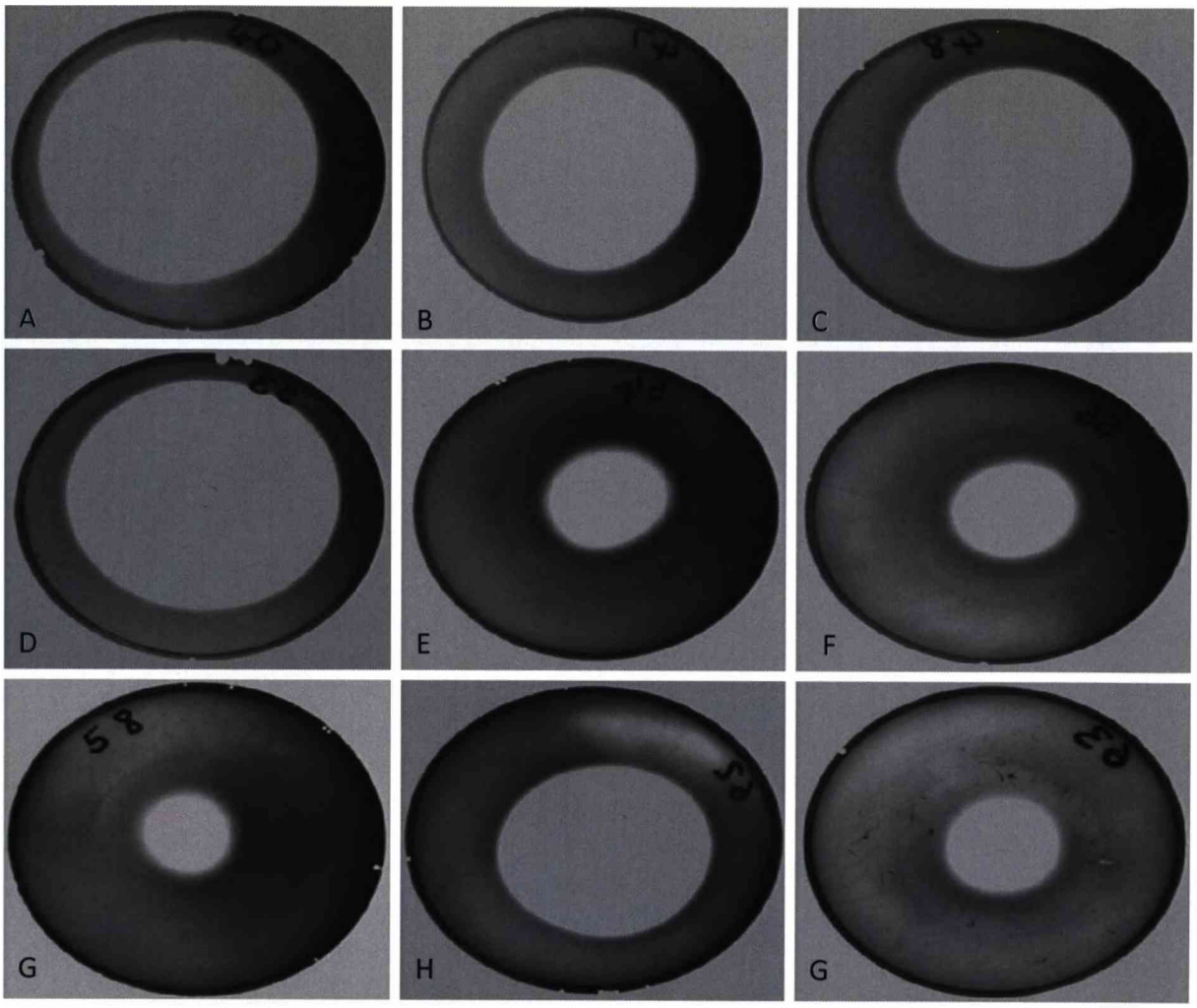


Figure5.1 Endoglucanase screening of Irish Sea bacterial isolates

0.2 μM filters (PALL) were aseptically placed on Marine broth 2216 agar plates that contained 0.1 % CMC. The filters were inoculated with the 26 Liverpool Bay isolates and incubated at room temperature for 4 days. Following incubation, filters were removed and the plates flooded with Congo red (0.1 %) for 30 min then washed 2 x 15 min with NaCl (1 M).Images A-G represent isolates 40, 47, 48, 53, 54, 56, 58, 62 and 63. Plates were viewed using a Syngenta imaging system using GeneSnap software. Zones of clearing in the middle of the plates can be seen. These zones are produced when the Congo red can no longer bind the β 1,4-glycosidic bond of the cellulose, when it has been hydrolysed by endoglucanase (however no lab strain such as *E.coli* were used to test this screening method).

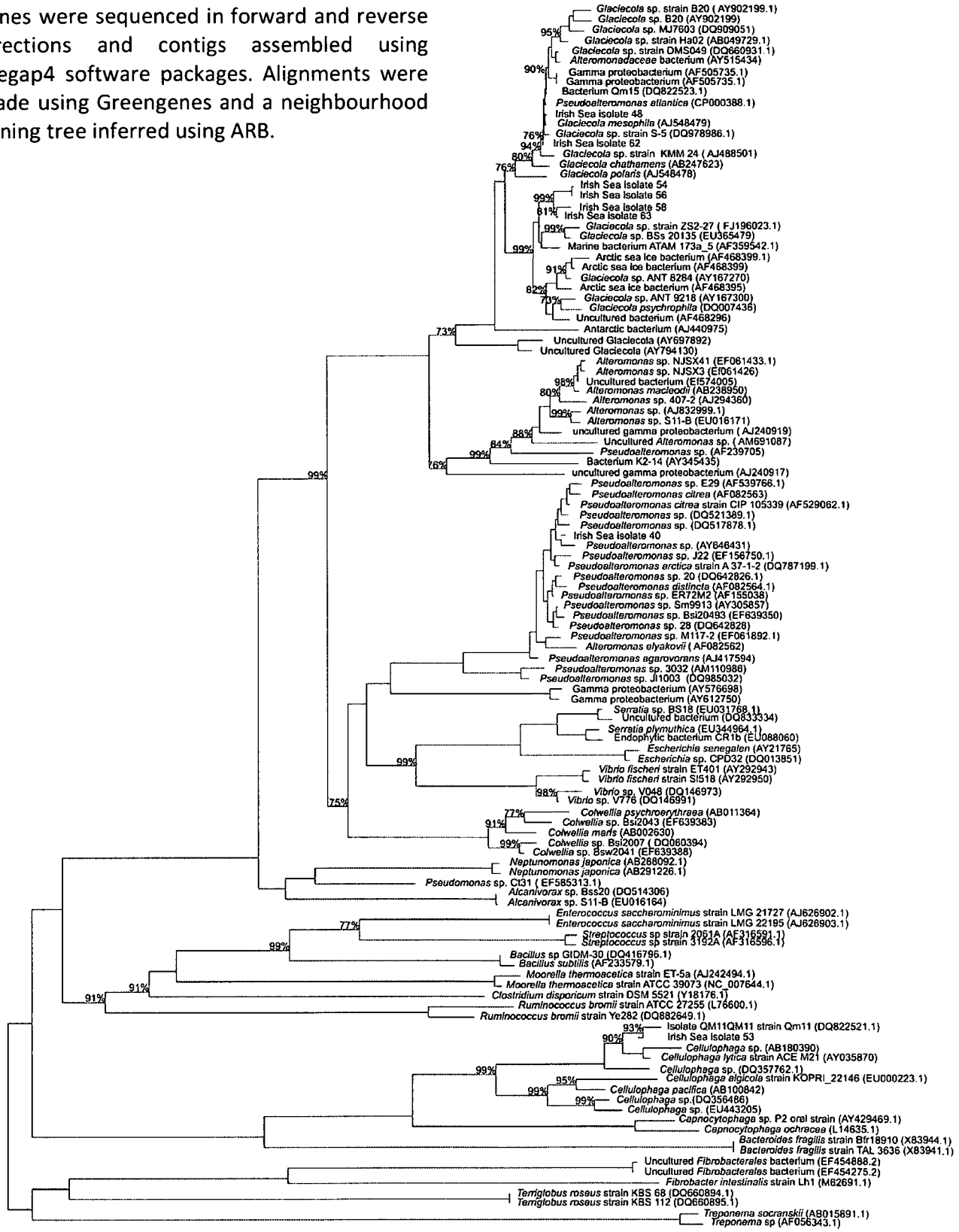
5.3.2 Phylogenetic analysis of 16S rRNA genes of Irish Sea bacterial isolates

General bacterial primers pA and pH' (Edwards *et al.*, 1989) were used to amplify almost the entire 16S rRNA gene (~1534 bp) of all nine endoglucanase positive isolates using Phusion® Taq polymerase (Finnzymes). The amplification products were excised from an agarose gel, purified and cloned. Plasmid DNA was extracted, sequenced and subjected to phylogenetic analysis. Assembled contigs were analysed by two chimera check packages, RDP Chimera Check (Cole *et al.*, 2005) and Pintail (Ashelford *et al.*, 2005). Both chimera detection packages identified an anomaly with isolate 47 indicating that the amplified 16S rRNA gene that was sequenced was chimeric. Although isolate 47 was ostensibly grown as a pure culture, the amplified product, once sequenced, contained two halves of a 16S rRNA gene from two different species of bacteria. One half of the sequence was closely matched to *Glaciecola* whilst the other was most similar to a sequence from *Bacillus thermoleovorans*. The positive control used in PCR amplification was of an amplified *Bacillus thermoleovorans* 16S rRNA gene suggesting contamination during sample preparation for PCR amplification.

Figure 5.2 illustrates the phylogenetic relationship between eight of the Irish Sea isolates and their closest matched sequences from the Greengenes database (isolate 47 was not included). An alignment was constructed in ARB using all eight Irish Sea 16S rRNA gene sequences (not including isolate 47) and closest matches retrieved from Greengenes and a neighbour-joining tree with 1000 samplings was constructed from the alignment. A maximum-likelihood tree was also produced using PhyML. Branching patterns of the maximum-likelihood tree was the same as that of the neighbour-joining tree. (data not shown). Isolate 40 was placed within the Psudoalteromonads; isolate 53 clustered within the *Cellulophaga* lineage; Four isolates, 54, 56, 58 and 63 formed a distinct cluster within the *Glaciecola* lineage, with a bootstrap value of 99 %.

Figure5.2 Neighbour-joining tree of Irish Sea bacterial isolates.

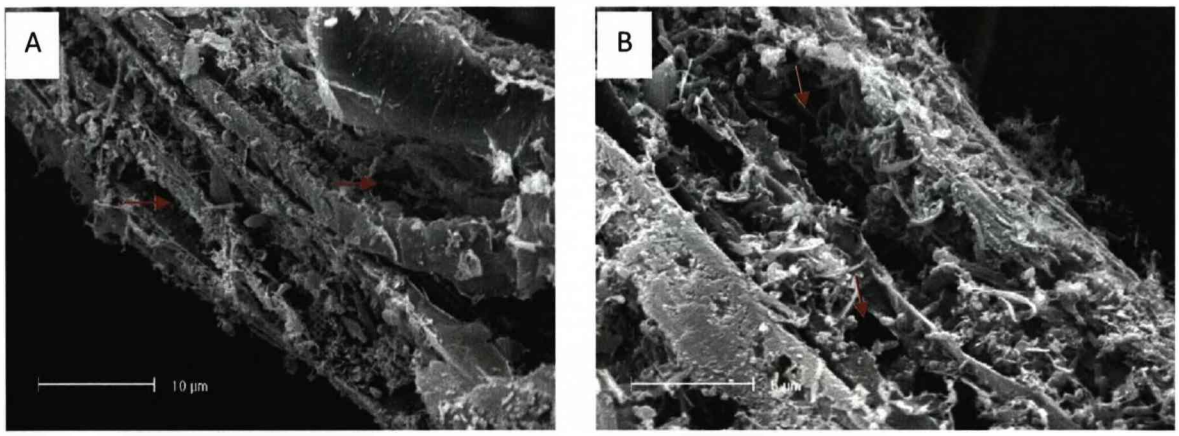
Irish Sea isolates highlighted in blue. 16S rRNA genes were sequenced in forward and reverse directions and contigs assembled using pregap4 software packages. Alignments were made using Greengenes and a neighbourhood joining tree inferred using ARB.



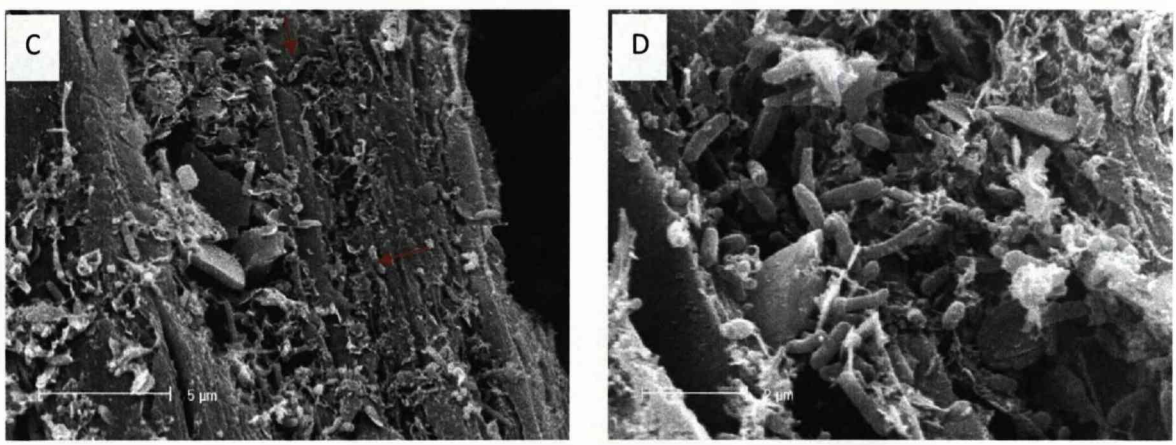
5.3.3 Scanning Electron Microscopy (SEM) of colonised cotton

The surface of colonised cellulose from Buoy B in the Irish Sea was visualised using SEM. The surface of cellulose not colonised has previously been visulaised also (McDonald, 2008). The cellulose bait was seen to be heavily colonised (Fig 5.3 A-J), and the morphology of the colonising microorganisms was predominantly rod shaped (G-J); cocci (G & J), filamentous (D-H) and spiral (J) morphologies could also be observed. Microorganisms were arranged in rows on the cellulose surface (I) and in a random manner (J), with pockets visible where degradation had occurred (D-J). In areas of heavy colonisation by a dense biofilm matrix, visible signs of degradation could be seen where the cotton surface had been eroded (A-J). Notably microorganisms appeared typically small (<1 μM) as is often common with isolates from the marine environment. Protuberances can be seen on the surface of some rod shaped cells (J). One explanation for the appearance of protuberances could be the presence of cellulosomes, which are known to exist on the surface of anaerobic bacteria for which the Gram positive organism *Clostridium thermocellum* has become the model organism (Bayer *et al.*, 1985). *Saccharophagus degradans*, the model marine aerobic polysaccharide degrading microorganism has also been shown to produce cell attached surface structures when cells were examined by Scanning Electron Microscopy following propagation in the presence of cellulose, (Ekborg *et al.*, 2005). A number of hydrolytic enzymes are thought to be consorted in such *S. degradans* complexes (Weiner *et al.*, 2008). It is possible that marine bacteria capable of colonising and degrading cellulose would in fact produce such structures to adhere themselves and their hydrolytic enzymes to the substrate surface, limiting loss of enzyme or hydrolysis products from the cell. It has been suggested by genome sequence analysis of *S. degradans* and *Teredinibacter turneae* (another marine cellulose degrading bacterium (Yang *et al.*, 2009)) that Gram negative bacteria use lipoproteins in the anchoring of carbohydrate active enzymes to the outer membrane, playing a similar role to that of cellulosomes that are present in Gram positive bacteria (Weiner *et al* 2008; Yang *et al.*, 2009).

Figure 5.3 Scanning Electron Microscopy of colonised cellulose baits from the Irish Sea Buoy B site

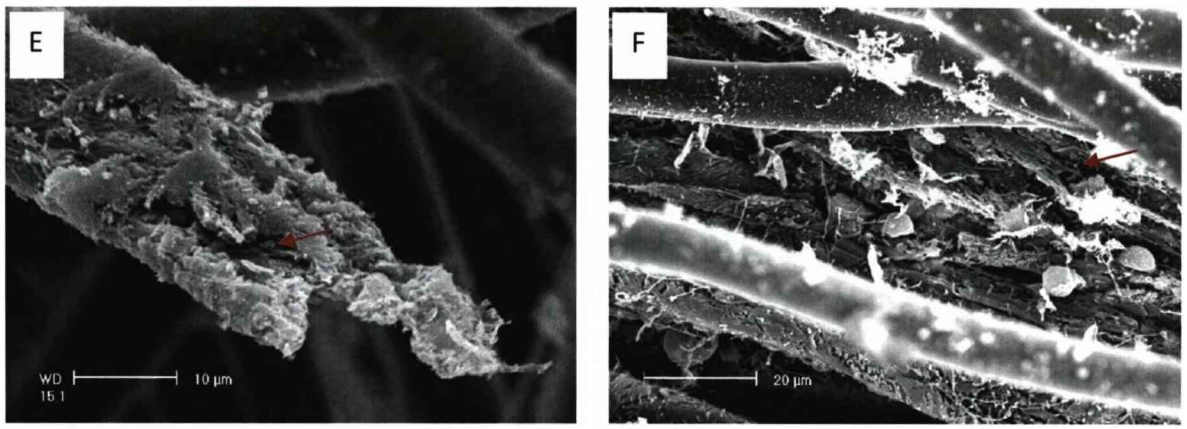


A-shows a heavily digested cellulose fibre colonised by a biofilm matrix. A number of hollow regions are visible within the cellulose fibre (shown by arrows). B- A closer view (bar = 5 µM) of the cellulose fibre shows heavily degraded eroded regions colonised by biofilm matrix (shown by arrows).

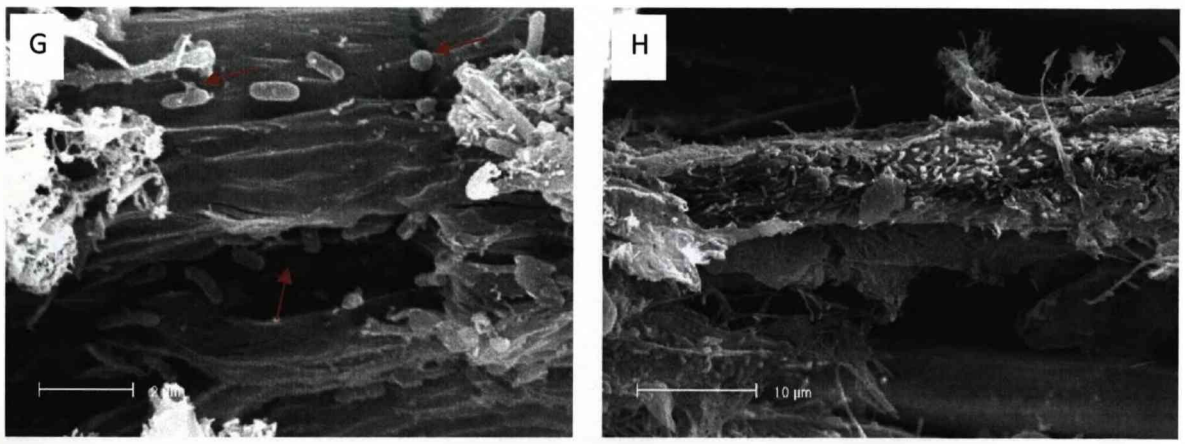


C-shows a cellulose fibre heavily colonised by a biofilm matrix. Rod shaped bacteria can clearly be seen sitting in grooves (arrow) providing visible evidence of degradation by cellulolytic bacteria. D-

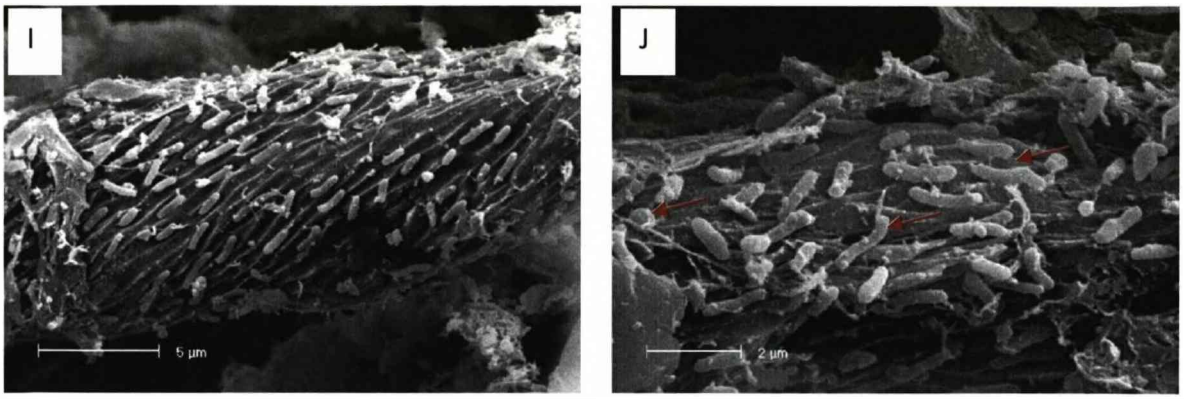
Shows a close image (bar = 2 μM) of a hollowed out region of a cellulose fibre. The region is abundant in rod shaped bacteria and filamentous material.



E-shows the end of a cellulose fibre which has been heavily colonised and shows clear signs of degradation characterised by hollowed out regions of the fibre. F-shows an overview of a number of cellulose fibres; areas of degradation can clearly be seen (shown by arrows).



G-A number of rod shaped bacteria are visible in the hollowed out region of a cellulose fibre (arrows). Cocci (shown by arrow) are also visible on the cellulose fibre surface. H- Partially digested cellulose fibre is heavily colonised by rod shaped bacteria and filamentous material.



I – shows the arrangement of rod shaped bacteria in rows around the cellulose surface. Grooves are visible where bacterial cells have formed pits on the cellulose surface. J-shows a closer view of a cellulose fibre where mainly rod shaped bacteria are arranged randomly and are clearly sitting in grooves on the surface. Some spirilla (arrow) and cocci (arrow) can also be seen. Protuberances are visible on the surface of a number of the cells on the cellulose surface.

5.4 Discussion

Enrichments whereby measures are used for the selective attachment and growth of functional bacterial groups of interest have previously been employed to mimic natural environments for the study of resident bacterial communities (Wery *et al.*, 2003; Takii *et al.*, 2008). Bacterial populations have been cultured from surface associations with marine eukaryotes known to be a source of cellulose such as those of colonising algae (Jensen *et al.*, 1996; Penesyan *et al.*, 2009). But in general the ecology of cellulose degradation by marine microorganisms has not been intensively studied. Here, *in situ* enrichment for the isolation of novel bacterial species colonising the insoluble polysaccharide cellulose has been applied (for the first time in the marine environment).

The genus *Glaciecola*, proposed to accommodate Gram-negative, aerobic, psychrophilic, pigmented and seawater-requiring bacteria was proposed by Bowman *et al.*, (1998) following the isolation of two species of bacteria (*G. pallidula* and *G. punicea*), isolated from sea-ice diatom assemblages at coastal regions of eastern Antarctica. Through 16S rRNA analysis it was demonstrated that these two species formed a novel lineage adjacent to the genus *Alteromonas,* in the γ-proteobacteria subclass. Subsequently, a number of bacterial species have been isolated and undergone 16S rRNA gene analysis and taxonomic characterisation resulting in a growing number of representatives of the genus *Glaciecola.* To date, nine novel species have been affiliated to this genus: *G. pallidula* and *G. punicea* (Bowman *et al.*, 1998), *G. mesophila* (Romanenko *et al.*, 2003), *G. polaris* (Van Trappen *et al.*, 2004), *G. nitratireducens* (Baik *et al.*, 2006), *G. psychrophila* (Zhang *et al.*, 2006), *G. chathamensis* (Matsuyama *et al.*, 2006), *G. agarilytica* (Yong *et al.*, 2007) and *G. lipolytica* (Chen *et al.*, 2009). Representatives of *Glaciecola* have subsequently been identified through environmental 16S rRNA gene analysis of DNA extracted from sea water (Junge *et al.*, 2001; Prabagaran *et al.*, 2006). Whilst the genus *Glaciecola* is located within the *Pseudoalteromonas, Alteromonas, Glaciecola* group of the γ-proteobacteria, unlike the pseudoalteromonads and alteromonads, cellulase production has yet to be described in *Glaciecola* species.

Glaciecola spp., although initially isolated from Antarctic Sea ice (Bowman *et al.*, 1998), have subsequently been isolated from a wide diversity of locations including the Arctic (Zhang *et al.*, 2006; Van Trappen *et al.*, 2004), surface sea water, from Korea and China (Baik *et al.*, 2006; Chen *et al.*, 2009) and in association with invertebrates; a strain of *G. mesophila* was isolated by Romanenko *et al.* (2003) from the internal liquor of *Halocynthia aurantium* (a marine ascidian) from coastal water in the Sea of Japan. Cellobiose utilisation is known in some *Glaciecola* species such as *G. mesophila* (Romanenko *et al.*, 2003), *G. polaris* (Van Trappen *et al.*, 2004) and *G. algarilyticus* (Yong *et al.*, 2007). Agar utilisation is also recognised in two species, *G. algarilyticus* (Yong *et al.*, 2007) and *G. mesophila* (Romanenko *et al.*, 2003) and Guo *et al.* (2009) reported the characterisation of a cold-active xylanase (XynA) belonging to GH family 10 of *Glaciecola mesophila* KMM 241. Of the Irish Sea isolates subjected to phylogenetic analysis here, six were found to be affiliated with the *Glaciecola*, and four of these clustered together in a distinct group. To date the *Glaciecola* are not known to contribute to cellulose degradation (cellulose degradation has not been tested in the literature) but now the colonisation of cellulose and the ability to hydrolyse CMC provides evidence for a role for members of this genus in the process of cellulose recycling in the marine environment. The complete genome sequence of a representative *Glaciecola* species is yet to be published, and this may confirm the presence of cellulolytic genes.

Members of the *Bacteroidetes*, previously described as the *Cytophaga-Flavobacteria-Bacteroides* (CFB) phylum have been described as being involved in the utilisation of DOM in oceans (Cottrell & Kirchman, 2000; Bauer *et al.*, 2006). Culture dependent and independent methods have been used to characterise *Cytophaga* spp. *Cytophaga*-like bacteria are unicellular, non-spore forming, Gram negative and represent an abundant group of bacteria in the ocean with cultured members being notably proficient in degradation of polysaccharides such as cellulose, chitin and pectin (Kirchman, 2002; Cottrell *et al.*, 2005). However, most of the details regarding marine polysaccharide degradation by *Cytophaga*-like bacteria are hypothesised (Cottrell *et al.*, 2005). A large

number of species affiliated to the genus *Cytophaga* have been reclassified into the *Cellulophaga* (Johansen *et al.*, 1999). The group is clearly diverse however little is known of their involvement in marine cellulose degradation. For example, there are no entries of glycosyl hydrolases from the genus *Cellulophaga* on the CAZy database.

Members of the genus *Pseudoalteromonas* are well documented as colonising marine surfaces, including higher marine organisms and eukaryotic algae. For example, *Pseudoalteromonas tunicata* was first isolated from a tunicate and later from an association with algae. *Pseudoalteromonas tunicata* has also been shown to increase expression of a mannose-sensitive haemagglutinin (MSHA)-like pilus in the presence of cellulose (Dalisay *et al.*, 2006) contained in marine algae. Cellulase production has been observed in a number of cultured strains of pseudoalteromonads (Xiong & Wen, 2004; Violot *etal.*, 2005) and psychrophilic cellulases have recently gained considerable interest for their potential in biotechnological uses (Violot *et al.*, 2005). It is not surprising therefore, that representatives of these groups of bacteria should colonise the cellulose baits, but it does provide further evidence for their involvement in degrative processes. What is surprising is the attachment of members of the genus *Glaciecola* exhibiting endoglucanase activity, as well as the more predictable representation of *Cellulophaga*. However as members of the *Glaciecola* are known to colonise algae and diatoms, this would suggest involvement in the biodegradation process. This very limited study on the identity of some endoglucanase positive isolates from *in situ* colonised cellulose does help to confirm that they have a role in marine cellulose degradation, and their importance is also supported by the representation of sequences from these organisms in the metagenome sequence dataset discussed in chapter 3.

Colonised cellulose bait from the buoy B site was heavily colonised by bacterial biofims. Throughout the bait as examined by SEM, areas of degradation were apparent. There is clearly a shortage of information on all bacteria involved in complex polysaccharide degradation in the marine environment, and while there is likely to be an ever increasing number of metagenomic studies, culturing marine bacteria using selective

measures such as the use of surface attached bacteria as the source material remains an effective approach to the study of marine bacterial communities.

5.5 Conclusions

- Twenty six bacterial strains were isolated from the biofilm colonising cotton cellulose bait from the Irish Sea.
- All strains were screened for endoglucanase activity using the CMC/Congo red staining method, and nine strains were positive.
- General 16S rRNA primers were used to amplify the 16S rRNA genes of all nine isolates. The amplified products were cloned sequenced.
- Using the 16S rRNA gene sequences, eight of the isolates were identified. Six of the isolates were members of the genus *Glaciecola*, of which four clustered together in a novel lineage. One isolate was clustered with the genus *Cellulophaga* and one isolate was found to cluster with Pseudoalteromonads.
- This is the first report of endoglucanase activity in members of the genus *Glaciecola.*
- Colonised string from the Irish Sea was examined by SEM and the biofilm community was shown to be dominated by rod shaped bacteria, surrounded by pits and hollowed out regions on the cellulose strands, showing *in situ* degradative activity.

The construction of a fosmid library using colonised polysaccharide bait DNA from Liverpool Bay was attempted however sufficient sized DNA could not be retrieved. Thus in the next chapter a pre-constructed library constructed from marine estuary sediment DNA was screened.

Chapter 6

Screening a fosmid metagenome library constructed from Colne Estuary sediment.

6.1 Introduction

Fosmids, like cosmids are hybrid plasmids containing a cohesive end (*cos)* site derived from λ phage that can be packaged into bacteriophage particles and used to build genomic libraries with inserts of ~30-50 kb (Collins & Hohn, 1978). Fosmids are maintained in *E.coli* in low copy number based on the F factor plasmid replication origin. Cosmids in contrast can be highly unstable and prone to deletions and recombination during maintenance and propagation as they are generally maintained in high copy number. Stability of inserts is essential when DNA inserts are large and it has been found that clones based on the Fosmid vector show a reduced frequency of detectable changes (Kim *et al*, 1992) providing an advantage for some DNA fragments that may be difficult to clone, for example if they express proteins that are toxic to a cell at high concentrations. Therefore, maintaining DNA in a low copy number fosmid allows for greater probability of DNA being successfully cloned and expressed, providing increased representation of the metagenome being examined (Kim *et al.*, 1992).

Fosmid libraries have been successfully constructed with DNA extracted from a number of microbial habitats including: soil (Kim *et al.*, 2008); seawater (Woebken *et al.*, 2007); marine sediment (Park *et al.,* 2008); activated sludge (Suenaga *et al.*, 2007); river estuary sediment (Meng *et al.*, 2008). Fosmid libraries are useful metagenomic tools that allow storage of large genomic fragments (~30-50kb) for long periods of time in the form of a stable clone library and can be screened by different techniques enabling investigation of various properties of the same microbial community.

Phylogenetic screening uses PCR to amplify and identify phylogenetic markers of interest such as 16S rRNA genes, present in clones across the library. This can be achieved

by general 16S rRNA gene amplification or selection for gene sequences specific to known prokaryote or eukaryote taxa (Woebken *et al.*, 2007). Measuring the phylogenetic diversity of a community and linking it to function in individual clones can also be achieved, providing that genes next to phylogenetic markers can be sequenced. 16S rRNA genes have also been identified using metagenome microarrays (MGA) (Park *et al.*, 2008) and Large insert Library FISH (LIL-FISH) (Leveau *et al.*, 2004) using the combination of gene expression/activity screening with phylogenetic screening.

Functional screening is a valuable tool for the detection of novel enzymes, however it is reliant on genes of interest being transcribed, translated and the resultant proteins being folded correctly in the heterologous host. Activity screening has previously been applied to screening fosmid libraries for the identification and characterisation of cellulases (Kim *et al.*, 2008) and extradiol dioxygenases (EDOs) (Suenaga *et al.*, 2007), using carboxymethyl cellulose and catechol as substrates for screening respectively. An N-acylhomoserine lactonase was also detected from a fosmid library constructed from soil DNA by screening the fosmid library for N-acylhomoserine lactone (NAHL) degradation (Riaz, 2008). Enrichment cultures have been employed in fosmid library construction, allowing selection of organisms capable of expressing the enzymes of interest. Neufeld *et al.* (2008) employed Stable Isotope probing (SIP) to select for metabolically active members of a marine surface water community. ^{13}C-labelled methanol was used as the substrate resulting in the labelling of biomass of organisms capable of metabolising the substrate. Combining this method with Multiple Displacement Amplification (MDA) which is a non-PCR based amplification technique (Dean *et al.*, 2002) a fosmid library was successfully constructed and screened for methanol dehydrogenase genes using PCR, allowing selection and identification of active prokaryotes within the community.

Fosmid libraries have great potential for ascertaining the phylogenetic and functional diversity of microbes which are unculturable and for which detailed characterisation would otherwise be impossible. This was exemplified when Meng *et al.* (2008) identified bacterio-chlorophyll in a member of the domain *Archaea*, a pigment

previously only identified in *Bacteria* and not hitherto present in any culturable Archaeal species. Fosmid libraries can provide access to information on the key members of microbial communities and their function within that community, providing an attractive prospect for the search and discovery of novel products for use in biotechnological processes.

6.1.2 Sample site

The River Colne estuary located on the east coast of the United Kingdom is a small, macro-tidal (3 to 5 m), hypernutrified, muddy estuary entering the North Sea at Brightlingsea, Essex (Nogales *et al.*,2002; Kondo *et al.*, 2007). There is a large amount of biogeochemical information on the Colne estuary, which has been subjected to a number of investigations including those focussed on methanogenesis (Purdy *et al.*, 2002), nitrogen cycling (Nogales *et al.*, 2002; Smith *et al.*, 2007) and sulphur cycling (Kondo *et al.*, 2007). A fosmid clone library was constructed from Colne estuary sediment by Dr Ashley Houlden, University of Sheffield. A copy of the library was provided in the form of 96-well cultures, containing 8 % DMSO for storage at -80 ^{o}C. The library was functionally screened for the presence of endoglucanase encoding genes.

6.1.3 Aims and objectives

The aims of this chapter were to screen a fosmid library of DNA from Colne Estuary sediment for endoglucanase genes by first developing a functional screening method for expressed cellulase activity. Endoglucanse positive clones were sequenced and annotated to locate possible ORF's. ORF's regarded as potentially expressing endoglucanase following *in silico* analysis were subject to sub cloning and further expression analysis to decipher the ORF/s responsible for the expression of endoglucanase.

6.2 Materials and methods

6.2.1 Bacterial Strains and plasmids

Escherichia coli (*E.coli*) strains were grown at 37°C in Luria-Bertani (LB) broth or agar supplemented with appropriate antibiotics. Antibiotics were added to media at final concentrations of 50 (kanamycin) or 12.5 (chloramphenicol) μg ml^{-1}.

6.2.2 Fosmid library construction

Sampling, DNA extraction and fosmid library construction was performed by Dr Ashley Houlden, University of Sheffield. Sediment samples (0-5 cm depth) were collected on the 25th October 2005 from the Colne estuary, U.K. (51°52.4 N, 0°55.5 E) at low tide. Bacteria were removed from sediment using blending and differential speed centrifugation and subsequently lysed for the extraction of DNA. High molecular weight DNA was extracted using an optimised indirect lysis method, modified from Gabor *et al.* (2003). DNA samples were visualized using pulse field gel electrophoresis (PFGE; 6 V cm^{-1}, 0.1-8 s pulse times, 14 h at 11°C , CHEF Mapper BioRad, USA) and a fosmid library constructed according to the Copy Control™ Fosmid Library Production Kit (EPICENTRE), using DNA size fractionated to 35-50 kb. Clones were picked and propagated in 96-well plates containing 100 μl Luria-Bertani broth with 12.5 μg ml^{-1} chloramphenicol (Sigma, UK) (as the vector carried a chloramphenicol resistance gene) at 37 °C. 50 μl of 75 % (v/v) glycerol were added to overnight cultures and clones were stored at – 80 °C.

6.2.3 Fosmid Library Screening

The fosmid library was supplied in 96-well culture plates, containing 8 % (v/v) DMSO for storage at -80 °C. Q-Trays (Genetix, New Milton, UK) containing Luria-Bertani supplemented with agar (1.8 %), CMC (0.1% w/v) and chloramphenicol (12.5 μg ml^{-1}) were used for screening the fosmid clone library for endoglucanase activity. Clones were replicated onto agar with a 96-pin replicator (6X96 on each Q-Tray). These were incubated at 37°C overnight with 24 h further incubation on the bench. Colonies were blotted off the agar using damp Whatman filter paper and the agar subsequently

stained by flooding with Congo red (0.1 %) for 30 min with shaking. The Congo red was removed and Q-trays were washed with NaCl (1 M) for 30 min, with the NaCl changed twice. Clones positive for endoglucanase were identified by the presence of a yellow zone around the clone, against a red background.

6.2.4 Sequence Analysis

GenBank database searches were carried out using various BLAST programs on the NCBI website (http://www.ncbi.nlm.nih.gov/BLAST/).

The ExPASy PEPTIDEMASS tool (http://www.expasy.ch/tools/peptide-mass-ref.html) was used for hypothetical protein digest using trypsin. Molecular weights and isoelectric points (pI) were calculated for ORFs using the ExPASy compute MW/pI tool (http://www.expasy.ch/tools/pi_tool.html)

6.2.5 Protein extraction using non-denaturing conditions

Luria-Bertani broth (10 ml) supplemented with corresponding antibiotic was inoculated with the clone of interest and incubated overnight at 37°C with shaking at 200 rpm. 50 µl of overnight culture was used to inoculate a fresh 10 ml aliquot of Luria-Bertani broth containing antibiotic in a universal tube and grown up to an OD_{600} of 0.6 at 37 °C. Cells were subsequently harvested by centrifugation at 5000 x g for 5 min and the pellet resuspended in 1 ml 0.1 M sodium phosphate buffer, pH 6.4. Cells were then sonicated 6 X 10 s, held on ice between treatments for at least 30 sec. Cell debris was removed by centrifugation for 10 min at 14,000 x g and the supernatant removed to a fresh 1.5 ml microfuge tube.

6.2.6 Zymogram analysis

CMC was incorporated into the separating portion of a SDS-PAGE gel to a final concentration of 0.1% w/v. Protein was quantified using the Sigma BCA method according to manufacturer's instructions and heated at 80 °C, 10 min in 1 X laemmli buffer (Laemmli, 1970). After separation of the proteins, the zymogram was washed in ddH_2O, 2 X 20 min at 4°C, then placed in 100 ml renaturation buffer, containing 0.1 M

sodium phosphate buffer, pH 6.4 with 0.1 % v/v Triton X-100, and incubated for 20 h at 37°C. The gel was subsequently stained with Congo red for 30 min and destained with NaCl (1M). Gels were examined for activity by the presence of zoned of clearing around protein bands. Images were captured using either the GE Healthcare Image scanner III or a Syngenta imaging system with GeneSnap software.

6.2.7 Fosmid sequencing and analysis

An overnight culture of the required fosmid clone was grown in 10 mL Luria-Bertani broth containing chloramphenicol. Cells were collected by centrifugation at 5000 x g, for 5 min and resuspended in 1 mL Luria-Bertani broth containing chloramphenicol and 8 % DMSO. The resuspended cell pellet was sent MWG Biotech who extracted fosmid DNA for sequencing. Open Reading Frames (ORFs) were predicted using GLIMMER (Gene Locator and Interpolated Markov ModelER) software (Salzberg *et al.*, 1998) by Dr Mike Cox. ORFs were analysed using the Artemis program and individual ORFs were then searched for matches using the BLAST algorithm at NCBI (http://www.ncbi.nlm.nih.gov).

6.2.8 Design and optimization of predicted GH ORF specific PCR primer sets

Fosmid clones were propagated in 10 ml Luria-Bertani broth supplemented with chloramphenicol and incubated overnight at 37°C. Cells were collected by centrifugation at 5000 x g and Fosmid DNA was purified using the fosmidMAX™ DNA purification kit (Epicentre® Biotechnologies) according to the manufacturer's instructions. PCR primers were designed for four ORFs and reactions to amplify the predicted ORFs were performed in 50 µL volumes containing: 0.2 mM each primer, 0.2 mM each dNTP, 1 x Phusion® HF buffer (Finnzymes), 1 U Phusion® High fidelity DNA Polymerase (Finnzymmes), approximately 20 ng of purified fosmid DNA and ddH_2O. PCR cycling conditions were as follows: 98°C for 30 s, 30 cycles of 98°C for 10 s, 30 s at an annealing temperature range for each primer set of 55.2°C-74.4°C and an extension of 72 °C for 20 s followed by a final extension of 72 °C for 8 min.

6.2.9 Cloning and expression of glycosyl hydrolases

PCR amplification products were excised from 1 % agarose gels using a sterile scalpel blade and purified using a Perfectprep® Gel Cleanup kit (Eppendorf) following the manufacturer's protocol. Each amplified PCR product preparation was ligated into the pET30c expression vector (Novagen) following digestion with the corresponding restriction enzymes (Table 6.3) under the control of the T7 promoter. Each standard ligation contained a 10:1 ratio of insert DNA to plasmid; 100 U T4 DNA Ligase (New England Biolabs); 2 µl 10 x T4 DNA Ligase reaction buffer (New England Biolabs), to a total reaction volume of 20 µl. This was incubated at 16 °C for 16 h and the ligase inactivated by heating at 70 °C for 10 min.

Ligated pET30c was transformed into *E.coli* BL21 (DE3) chemically competent cells (Bioline). Ligation products were added to 50 µl cells and incubated on ice for 30 min. Cells were subsequently heat-shocked at 42 °C for 45 sec and then replaced on ice for 2 min. The transformation reaction was diluted to 1 ml with SOC medium (2% tryptone, 0.5% yeast extract, 0.4% glucose, 10 mM NaCl, 2.5 mM $MgCl_2$ & 10 mM $MgSO_4$) and incubated at 37°C for 60 min with shaking at 200 rpm. Transformed cells were spread plated onto LB agar containing 50 µg ml^{-1} kanamycin. After overnight incubation at 37°C, colonies were picked and streaked onto fresh LB agar/kanamycin plates and subjected to PCR amplification using the primers designed to each ORF to ensure presence of insert.

6.2.10 Agarose gel electrophoresis.

Agarose gel electrophoresis was used to visualise the products of DNA extraction and PCR amplifications.1 % (w/v) agarose was added 1x TAE (Tris-acetate EDTA) (50x TAE stock: 2M Tris; 57.1 ml l^{-1} glacial acetic acid; 0.05M EDTA; adjusted to pH 8.0) and heated to approximately 50 °C. Ethidium bromide was added to a concentration of 0.5 ng ml^{-1}, and the mixture poured into a gel-forming tray and allowed to set for 30 min. The gel was then immersed in a volume of 1 x TAE buffer sufficient to cover the wells and electrodes. DNA samples were mixed with 6x loading dye (Sambrook & Russell, 2001) and run alongside molecular weight marker Hyperladder I (Bioline). The gels were run at 60 mA for 1-1.5 h and were viewed by

UV-transillumination and the presence of DNA recorded using a Syngenta imaging system with GeneSnap software.

6.2.11 Induction of ORF 10 clone

10 ml LB broth supplemented with kanamycin (50 mg ml^{-1}) was inoculated with the ORF 10 clone and incubated overnight at 37 ^{0}C with shaking at 200 rpm. 100 ml LB broth (50 mg ml^{-1} kanamycin) was inoculated with 1 ml of overnight culture and incubated at 37 ^{0}C with shaking at 200 rpm until an OD_{600} of 0.6 was recorded. 0.5 ml culture was removed and added to a microfuge tube and the cells collected by centrifugation at 5000 x g for 10 min. The supernatant was discarded and 80 µl Laemmli buffer (1 x) added to the pellet and the mixture heated at 95 ^{0}C for 3 min. 10 µl of each lysate was separated by 1D SDS-PAGE (4.2.3.2). To the remaining culture, IPTG (1 mM) was added and incubated at 37 ^{0}C with shaking at 200 rpm. After 1 h, 2 h and 3 h incubation periods, 0.5 ml was removed and treated as above. In addition BL21 (DE3) cells were transformed with pET30c vector without the ORF10 insert and treated as above to act as a control.

6.2.12 Protein extraction under denaturing conditions

A 100 ml culture of ORF 10 clone was incubated at 37 ^{0}C with shaking at 200 rpm until an OD_{600} of 0.6 was recorded. 0.5 ml was removed and centrifuged at 5000 x g for 10 min. The supernatant was discarded and the pellet resuspended in 80 µl Laemmli buffer (1x) and heated at 95^{0}C, for 3 min, then centrifuged at 5000 x g for 10 min and the supernatant removed to a clean microfuge tube. The remaining culture was induced with 1 mM IPTG and incubated for a further 3 h. 3 ml was removed and centrifuged at 5000 x g, for 10 min and the pellet resuspended in 480 µl PBS. The resuspended cells were sonicated 6 x 10 s intermittently on ice and aliquoted into 5 x 80 µl in fresh microfuge tubes and the cell debris collected by centrifugation at 5000 x g, for 10 min. For four tubes, the supernatant was discarded and the pellets resuspended in 2 M, 4 M, 6 M or 8 M urea (80 µl). 10 µl of this was added to Laemmli

buffer. Ten µl of the remaining aliquot was added to Laemmli buffer as above and all were heated at 95^0C for 3 min. Proteins were separated by SDS-PAGE.

6.2.13 Purification of His-tag fusion recombinant protein (ORF 10)

A 1 L culture of the ORF 10 clone was propagated until an OD_{600} of 0.6 was recorded, 1 mM IPTG added and the culture incubated for a further 3 h. Cells were collected by centrifuging at 14,000 x g for 10 min and resuspended in 50 mM Tris-HCl pH 8.2. Cells were then sonicated in 1 ml aliquots for 6 x 10 secs intermittently on ice. Sonicates were then pooled and centrifuged at 14,000 x g for 10 min. The pellet was washed with 10 ml 50 mM Tris-HCl containing 1 M urea and 1% Triton X-100. The washed cell debris was then pelleted by centrifugation at 14,000 x g for 10 min and the pellet resuspended in 50 mM Tris pH 8.2 containing 8 M urea and DTT (50 mM), and centrifuged at 14,000 x g for 20 min and the supernatant removed. A His-Trap™ HP column (GE healthcare) was pre-equilibrated with 50 mM Tris-HCl pH 8.2 containing 8 M urea and the supernatant was then loaded at a flow rate of 1.0 ml min^{-1}. Following a wash with 50 mM Tris-HCl, pH 8.2, 8 M urea the bound proteins were eluted in steps of 20 mM imidazole, 100 mM imidazole and 250 mM imidazole in Tris-HCl pH 8.2, 8 M urea. Elution was monitored at 280 nm and 0.5 ml fractions were collected.

6.2.14 MALDI-ToF Analysis

The protein band of interest was manually excised from a polyacrylamide gel and in-gel digested with trypsin (4.2.5.2). Extracted peptides were analysed using a saturated solution of alpha-cyano-4 hydroxycinnaminic acid in 50% acetonitrile/0.1% trifloroacetioc acid performed using a MALDI-Tof instrument (Waters-Micromass). Peptides were selected in the mass range of 1000 – 4000 Da.

6.3 Results

6.3.1 Fosmid library screening

A total of 7104 fosmid clones were screened using a functional screening approach for endoglucanase activity. One clone- 01-102030 C4, termed hereafter as clone C4 was identified as having endoglucanase activity on LB agar supplemented with CMC (Figure 6.1). The clone was subsequently propagated and stored at -80°C supplemented with 8% DMSO for further investigation. Screening was carried out in the absence of an inducing agent under single vector copy conditions and endoglucanase only detected therefore if the gene was transcribed, translated and folded correctly in the heterologous *E.coli* host. Additionally detection depends on the cellular location of the enzyme, relying on enough active enzyme being released by host cells by leakage.

6.3.2 Zymogram Analysis of Fosmid clone C4

Endoglucanase activity of clone C4 was checked by zymogram analysis. Two 10 ml cultures of clone C4 were propagated and one culture was induced using the Copy Control™ Induction Solution (Epicentre® Biotechnologies) according to the manufacturer's instructions with scaling to a 10 ml culture. Cells from both cultures were subsequently harvested by centrifugation at 5000 x g for 5 min. Additionally the Epi300 *E. coli* (host strain used in fosmid library construction) strain was propagated in a 10 ml culture. Total protein extracts were produced, 10 µl of each separated by 1D SDS-PAGE and gels developed as a zymogram. The presence of 5 bands possessing endoglucanase activity in the C4 induced and uninduced cultures (Figure 6.2) were observed. The EPI300 *E.coli* protein extract also subjected to zymogram analysis, did not show endoglucanase activity, providing evidence that endoglucanase activity was a result of fosmid cloned genetic material.

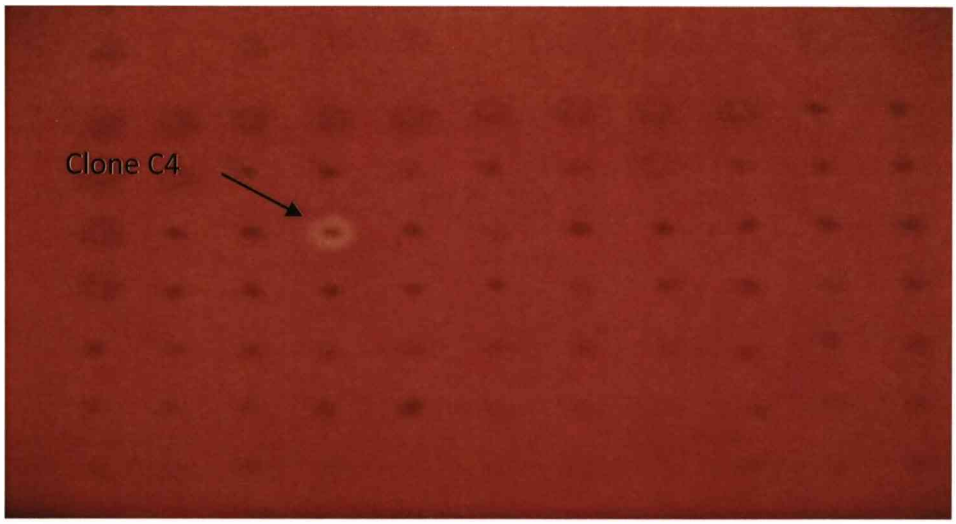


Figure 6.1 Fosmid library screening for endoglucanase activity.

Q-Trays containing Luria-Bertani supplemented with agar (1.8 % w/v), CMC (0.1% w/v) and chloramphenicol (12.5 μg ml^{-1}) were used for screening the fosmid clone library for endoglucanase activity. Clones were inoculated onto the agar with a 96-pin replicator. Following growth Q-Trays were stained with Congo red (0.1 % w/v) and washed with NaCl (1 M), changing the NaCl twice. The endoglucanase positive clone is identified by the yellow clearing zone against a red background (arrow). The clearing is a result of the Congo red no longer being able to bind the β 1, 4-bond of the cellulose polymer following bond hydrolysis from endoglucanase.

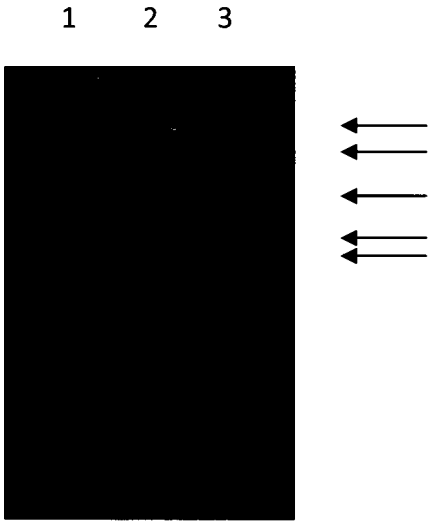


Figure 6.2 Zymogram analysis to detect endoglucanase proteins in cell lysates

Cell lysates of lane **1** EPI300 *E.coli*, lane **2** clone C4 and lane **3** induced clone C4 were prepared by sonication and proteins fractionated on a CMC zymogam. The proteins were washed 2 x 20 min in ddH_2O and incubated in 100 ml 0.1 M sodium phosphate buffer, pH 6.4 with 0.1 % v/v Triton X-100, and incubated for 20 h at 37°C. The gel was stained with Congo red and destained with 1 M NaCl to detect endoglucanase activity. Molecular weight markers were not used therefore molecular weight of active protein could not be estimated.

6.3.4 Fosmid assembly and annotation

The C4 fosmid clone, shown to contain the genetic information leading to expression of endoglucanase activity in *E.coli*, was sequenced by MWG biotech. Although 8-fold coverage was achieved, the sequence when assembled (by Dr Mike Cox) contained one gap resulting in two contigs. The overall size of the combined contigs was 38.7 Kb. A total of 29 ORFs were predicted using GLIMMER, comprising 10 ORFs covering 16880 bp of DNA (Table. 6.1) and 19 ORFs representing 21819 b of DNA (Table 6.2) for contigs one and two respectively. The translated sequences were produced using the Artemis program (Rutherford *et al.*, 2000) and used as query sequences against the blastp algorithm against sequences in the NCBI non-redundant database.

Table 6.1 and 6.2 show top blastp hits (December 2008) of all predicted ORFs from the two sequenced contigs of clone C4. A cluster of three ORFs located in tandem on contig two were highlighted as ORFs that potentially encoded endoglucanase, with closest hits inferring glycosyl hydrolases from families 5, 8 and 6 for ORFs 9, 10 and 11 respectively (Table 6.2). ORF 13 was also of interest with closest matches for protease and xylanase containing two CBM family 3 domains (Table 6.2).

Table 6.1 Overview of Fosmid clone C4- Contig 1 compared to the non-redundant (nr) protein database at GenBank

ORF	*Nucleotide range*	*Strand*	*G-C content (mol %)*	*Protein size (aa)*	*Best blast hit (e-value)*	**Conserved Domains**	*source*	*Identity (%) (overlapped aa)*
2	323-2881	+	53	853	XRE Family transcriptional regulator 4e-75	*-Helix-turn-helix (HTX_XRE) family* *-COG 3903 (predicted ATPase)* *-Tetratricopeptide (TPR)domain*	*Herpatosiphon aurantiacus*	30 (237/782)
3	2855-4057	-	53.4	401	Hypothetical protein 2e-33	*-No conserved domains*	*Nostoc punctiforme*	29 (105/374)
4	4486-6090	-	59.1	535	Carboxyl Transferase 2e-176	*-Acetyl-CoA Carboxylase (AccD)*	*Comamonas testosteroni*	57 (311/537)
5	6143-6616	-	48.5	158	Hypothetical protein 2e-19	*-No Conserved Domain*	*Rhodopirellula baltica*	40 (45/122)
6	6746-8296	-	55	517	Hypothetical protein 2e-114	-Amino Oxidase	*Burkholderia ubonensis*	42 (234/550)
7	8756-9292	-	45	179	Phage related protein 0.20	-No conserved domain	*Bacillus thuringiensis*	26 (42/158)
8	9317-10234	-	46.6	306	Hypothetical protein 3e-15	-DUF75 (protein of unknown function)	*Mycobacterium leprae*	25 (73/291)
9	11920-12882	-	56.4	321	Hypothetical protein 7e-42	-alkPPc superfamily (Alkaline - phosphatase homologues-linked with carbohydrate metabolism and transport)	*Desulfibacterium hafnience*	35 (96/270)
10	13427-14764	-	50	446	Major facilitator superfamily permease 7e-117	-Sugar_tr superfamily	*Roseoflexus castenholzii*	54 (245/449)
11	14877-16505	-	52.54	543	Cell envelope-related transcriptional attenuator 2e-33	-Lyt R_cpsA_psr superfamily	*Thermosinus carboxydivorans*	35 (85/240)

Table 6.2 Overview of Fosmid clone C4- Contig 2.compared to the non-redundant (nr) protein database at GenBank

ORF	Nucleotide range	Strand	G-C content (mol %)	Protein size (aa)	Best blast hit (e-value)	Conserved domains	source	Identity (overlapped aa)
1	395-1366	-	50.2	324	Translation elongation factor 7e-45	-Ef_TS superfamily	*Clostridium butyricum*	35 (112/313)
2	1405-1650	+	44.3	82	No significant Hits	No conserved domains		
3	1780-1950	-	44.4	57	No significant Hits	No conserved domains		
4	1867-2058	+	46.4	64	No significant Hits	No conserved domains		
5	1961-2323	-	48.5	121	Hypothetical conserved protein 9e-37	-DUF82 superfamily (no known function)	*Synechococcus sp*	65 (78/120)
6	2310-2543	-	44	78	Hypothetical protein 3e-24	DUF433 superfamily (no known function)	*Roseiflexus sp*	70 (55/78)
7	2626-3852	+	53.4	73	Putative ubiquitone biosynthesis protein 1.4	No conserved domain	*Acidovorans avense*	27 (15/54)
8	2602-2820	-	52.1	409	SoxB-like sarcosine oxidase, β subunit 2e-64	Pyr_redoxsuperfamily	*Geobacillus thermodentrificans*	36 (149/404)
9	3942-6287	-	51.4	782	Glycosyl Hydrolase family 6 protein 6e-163	-Glycosyl Hydrolase 6 superfamily -Cellulose Binding domain superfamily 2 (x2)	*Herpetosiphon aurantiacus*	62 (284/453)
10	6312-8780	-	52.5	823	Cellulose Binding family 2 protein 4e-164	-Cellulose binding domain superfamily 2 (x2) -Glycosyl hydrolase family 8 superfamily	*Herpetosiphon aurantiacus*	54 (317/580)
11	8934-11417	-	51.7	828	Endo-1,4-β-glucanase 0.0	-Carbohydrate binding super family 2 (x2) -Cellulase superfamily	*Reinekea sp*	56 (332/592)

12	11461-11703	+	49.8	81	No significant Hits			
13	12150-14408	-	46.7	753	SBA family peptidase 3e-40	-Carbohydrate Binding domain family 3 (x2) -PA Superfamily (protease associated domain)	*Synechococcus sp*	36 (129/355)
14	14338-17235	-	44.4	966	TPR repeat-containing protein 8e-119	-TPR (Tetracopeptide repeat domain) -COG4995 superfamily(uncharacterised conserved protein)	*Solibacter usitatus*	31 (297/942)
15	17235-17846	-	40.7	204	Hypothetical protein 0.003	No conserved domains	*Bacillus pumilus*	36 (29/80)
16	18074-18634	-	50.26	187	ECF subfamily RNA polymerase sigma-24 factor 6e-26	-Sigma 70_r2 Superfamily	*Solibacter usitatus*	34 (62/180)
17	18801-19409	-	47	203	McbG-like protein	-Pentapeptide superfamily	*Xanthomonas albilineans*	44 (85/193)
18	19405-20211	-	54.4	269	Hypothetical protein 5e-85	-AdoMet_MTases	*Gemmata obscuriglobus*	61 (157/255)
19	20329-21654	-	43.4	442	PAS/PAC sensor hybrid histidine kinase 1e-104	-HisKa superfamily -HATPase_c superfamily -Rec superfamily	*Pelobacter propionicus*	44 (201/452)

6.3.5 *In silico* analysis of predicted ORFs

All *in silico* analysis undertaken on four candidate ORFs containing GH domains or CBM domains was performed using the Pfam database v. 22 (http://pfam.sanger.ac.uk/) and the carbohydrate active enzyme (CAZy) web resource (http://www.cazy.org/). Searches were performed using predicted translated sequences obtained from the Artemis program (Rutherford *et al.*, 2000). Schematic diagrams were drawn representing the domain architecture resulting from information obtained from these analyses (Figure 6.3).

ORF 9 (Fig 6.3) is 2345 bp in length, coding for 782 aa with a predicted MW of 83 KDa and pI of 3.99. It has 62% identity to a protein produced by *Herpatosiphon aurantiacus*. From the Pfam database, the protein is predicted to consist of one glycosyl hydrolase (GH) 6 catalytic domain at the C-terminal, starting at residue 368 and ending at residue 731; the predicted active site is at residues 448, 497 and 717. There are two family 2 carbohydrate binding modules from residue 49 to 152, 103 aa in length. The second CBM 2 starts at residue 207 and finishes at residue 310, also 103 aa in length. Additionally the 2 CBMs are flanked by two linker modules of tetracopeptide repeats (XPTX). These are found in many proteins but their function is unclear. The first is 34 aa in length and the second 35aa.

Sequence analysis shows that ORF 10 (Fig 6.3) is 2468bp in length encoding a protein of 823 aa with a predicted molecular weight (MW) of 87 KDa and a predicted pI of 4.6. The amino acid sequence exhibits 54% identity (overlapping aa) with a protein produced by *Herpatosiphon aurantiacus*, which has a family 8 glycosyl hydrolase and one family 2 carbohydrate binding module. Through subsequent Pfam analysis ORF 10, was found to have significant sequence homology to three domains the first of a glycosyl hydrolase family 8 (GH8) catalytic domain and with two family 2 carbohydrate binding modules present at the at the C- terminus. The predicted architecture is novel because there is no known representative of a GH8 with two family 2 carbohydrate binding modules in the Pfam architecture database. The crystal structure of a family 8 GH has been determined for Cel A from *Clostridium thermocellum* (Alzari *et al.*, 1996). It is thought that family 8 GH domains fold into a

regular $(\alpha/\alpha)_6$ barrel formed by six inner and six outer α helices, with a highly conserved residue Glu95 having been assigned as the proton donor.

The predicted GH domain begins at residue 60 and ends at residue 422, being 362 aa in length. The first CBM starts at residue 464 and ends at residue 567 being 103 aa in length while the second CBM starts at 722 and ends at residue 822 being 100 aa in length (Pfam). The GH8 family comprises enzymes with several known functions including chitosanase, cellulase, and xylanase but primarily exhibit endoglucanse activity. They are known to act via an inverting mechanism with a predicted aspartate acting as the nucleophile and a glutamic acid as the proton donor in the catalytic reaction.

ORF 11 (Figure 6.3) is 2483 bp in length coding for an 828 aa protein with a predicted MW of 89 KDa. It has 56 % identity (overlapping aa) to a protein produced by *Reinekea* sp. When analysed using pfam for conserved protein family domains, the protein showed conserved domains for one GH 5 (cellulase superfamily) domain starting at residue 401 and ending at residue 789, with an active site predicted at residues 580 and 730. Two family 2 carbohydrate binding modules were also detected, the first starting at residue 65 and ending at 169 and the second starting at residue 235 and ending at residue 337 (104 and 102 aa in length respectively). Family 5, also known as the cellulase superfamily is one of the largest groups of glycosyl hydrolase proteins, having a variety of enzyme functions such as chitosanase, Endo-1,4 β-xylanase and cellulase.

ORF 13 (Fig 6.3) is 2258 bp in length coding for a 753 aa protein, with a predicted MW of 82 KDa. It has 36% identity (overlapping aa) to a protein produced by *Synechococcus* sp. The protein is predicted to be multidomain in composition, consisting of a peptidase catalytic domain and two family 3 carbohydrate binding modules. Unusually, none of the known peptidase architectures exist in combination with a carbohydrate binding module. However, there are examples of related catalytic domains that exist with dockerin modules and fibronectin domains which are commonly found with cellulases and chitinases (Kataeva *et al.*, 2002). The Pfam predicted subtilase family peptidase begins at residue 152 and finishes at 382, with the

predicted active site predicted to be residues 177, 225 and 368. The catalytic domain is followed by two CBM family 3 proteins, the first of which begins at residue 435 and finishes at residue 516, the second begins at residue 612 and ends at 689 (81 and 77 aa in length respectively). These are known to have cellulose binding activity but CBM domains belonging to family 3 are usually ~150 aa in length.

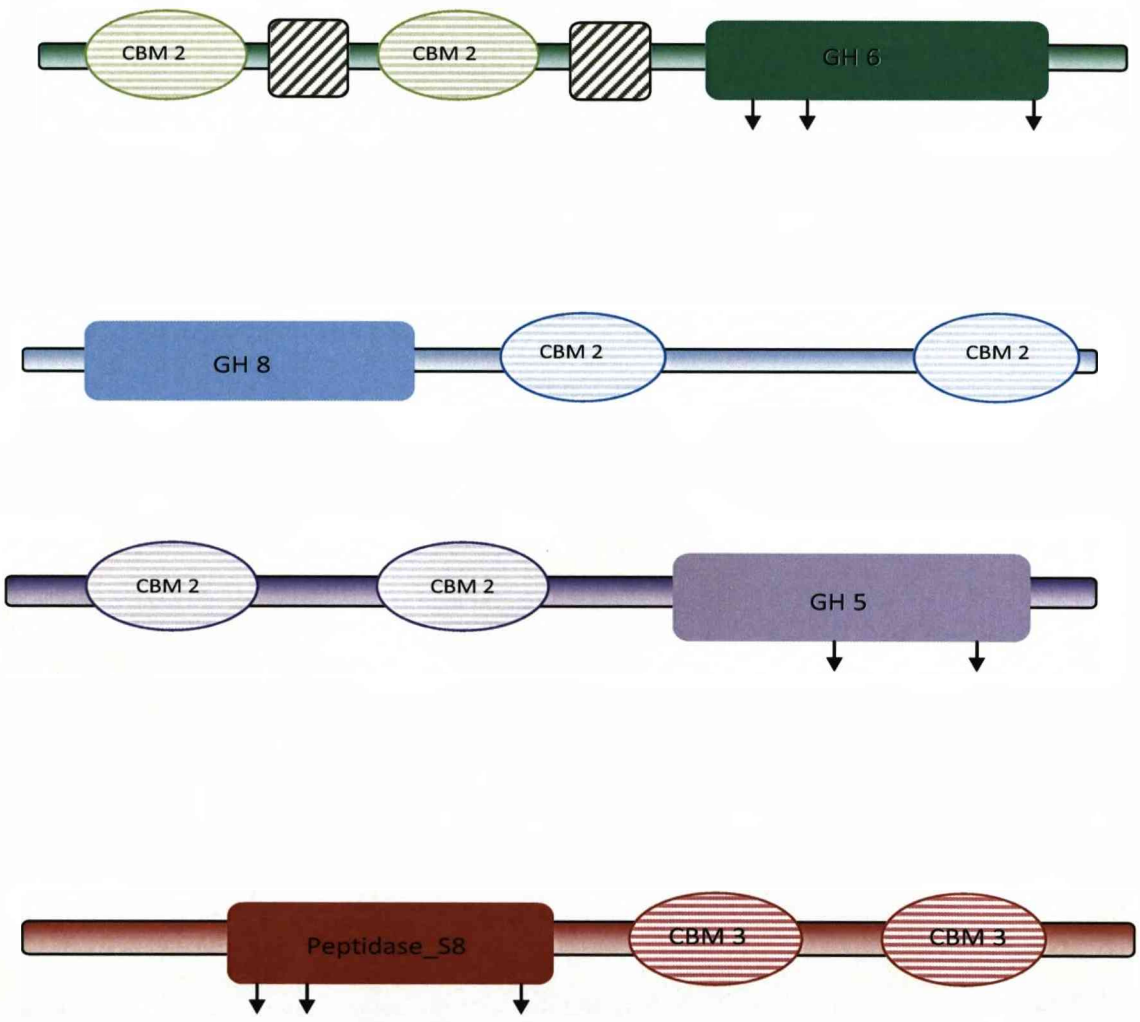


Figure 6.3 Predicted ORF protein domain architechture

Protein domain architecture as predicted according to the Pfam protein family database (http://pfam.sanger.ac.uk/) **A**, **B**, **C** and **D** represent ORF 9, 10, 11 and 13 predicted protein sequences respectively, resulting from the annotation of contig two of the C4 fosmid clone. Oblong shapes represent catalytic domains; oval shapes represent carbohydrate binding domains and the striped square boxes of **A** represent the two linker modules of tetracopeptide repeats (XPTX). Arrows indicate amino acids predicted (Pfam) to be involved in active sites.

6.3.6 Cloning ORFs of interest

The sub cloning of individual genes enabled the identification and confirmation of that or those ORFS responsible for the endoglucanase activity expressed by the fosmid clone. Primers were designed to amplify the four ORFs suspected to express endoglucanase activity based on the predicted ORF sequences 9, 10, 11 and 13 of contig two from the fosmid annotation, whilst introducing restriction sites into the cloned fragments (Table 6.3). The annealing temperatures for the PCR reactions were optimised using a temperature gradient and results showed that optimum annealing temperatures of 66.1°C, 63.6°C, 58.7°C and 60.9°C for ORF 9, 10, 11 and 13 respectively were most suitable and therefore used in the amplification of reactions ORF 9, ORF10, ORF11 and ORF13 (Figure 6.4). Clone C4 fosmid DNA was purified using the Fosmid MAX™ DNA purification kit (Epicentre® Biotechnologies), quantified by Nanodrop and 20 ng DNA used as template in each reaction for amplification of the four ORFs (figure 6.5). ORF 9, 10 and 11 primer sets amplified DNA fragments of the expected size ~2.3 kb, 2.5 kb, 2.5 kb (figure 6.5). However ORF 13 amplified DNA of slightly a smaller size than would be expected this might be due to issues with temperature optimisation as two bands are seen following amplification (Figure 6.4).

The amplified DNA was extracted from agarose gel slices using the Perfectprep® Gel Cleanup kit (Eppendorf), digested using corresponding restriction enzymes. ORF 9 and 10 amplified DNA fragments were digested with *Nco* I and *Eco*RI and ligated into the pET30c plasmid which was cut using the same restriction enzymes. The ORF 13 amplified DNA was digested with *Pci*I and *Eco*RI and ligated into the pET30c plasmid which had been cut with *Nco*I and *EcoR*I. The *NcoI/ Pci*I restriction site is no longer present with ligation producing a *Fat*I restriction site. ORF 11 was digested with *Bsp*HI and *Sac*I and ligated into pET30c plasmid digested with *Nco*I and *Sac*I (figure 6.6). The *NcoI/Bsp*HI restriction site is no longer present with ligation producing a *Fat I* restriction site. All primers would have been designed to include *NcoI* and *EcoRI* restriction sites into the cloned fragments; however the sequence analysis using pDRAW (http://www.acaclone.com/) located the presence of restriction sites of one or both within the ORF sequences and alternative enzymes with compatible cohesive

ends were used for ORF 11 and ORF 13 DNA (Figure 6.6). All restriction enzymes were obtained from New England Biolabs.

The pET30c (Novagen) vector which contained a gene for kanamycin resistance was chosen to produce a recombinant protein with an N-terminal 6-His-tag at a high level, following induction of the tightly controlled T7-promoter. The subsequent constructs were individually transformed into *E.coli* BL21 (DE3) cells (Bioline). Clones referred to as ORF9, ORF10, ORF11 and ORF13 were picked, propagated and subjected to PCR amplification with the ORF specific primers (Table 6.3) to ensure presence of the insert. Plasmid DNA was extracted and inserts were subjected to forward sequencing using the T7 specific primers at GATC biotech (http://www.gatc-biotech.com/). ~1 kb of sequenced DNA produced 100 % sequence matches to each of the predicted ORF sequences from the fosmid annotation.

Table 6.3 Primers designed to amplify selected ORFs

PRIMER	SEQUENCE	RESTRICTION ENZYME	ANNEALING TEMPERATURE
Orf 9F	CAGGGCCCATGGGTAACAGAAATTTGCG	*Nco*I	66.1°C
Orf 9R	CTCGAATTCTCAAGCGGAGGATAAGC	*Eco*RI	"
Orf 10F	CAGGCCCCATGGAAACGGAACAAGAA	*Nco*I	63.6°C
Orf 10R	CTCGAATTCTTAAACGTGGTAGGTGC	*Eco*RI	"
Orf 11F	CCGGTCATGACAGTTCAGGATG	*Bsp*HI	58.7°C
Orf 11R	CGCGAGCTCTCAACAAACTGGT	*Sac*I	"
Orf 13F	GCGCACATGTCGATGGAGCATAAAACAGC	*Pci*I	60.9°C
Orf 13R	CTCGAATTCTTAAGAATGAGTGATCG	*Eco*RI	"

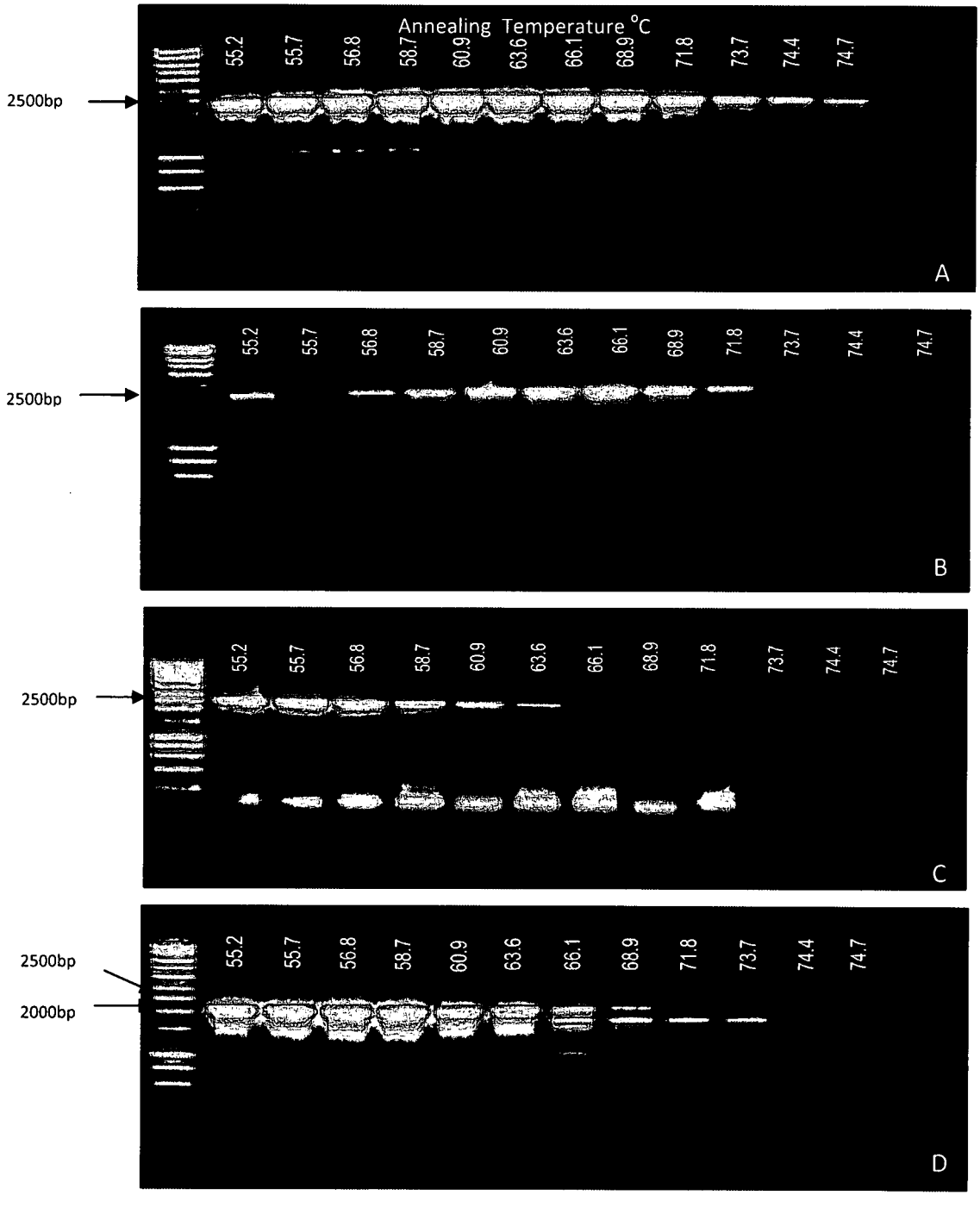


Figure 6.4 Determining the optimum annealing temperature for predicted ORF primer sets using Phusion High fidelity DNA polymerase.

Purified fosmid DNA was used as a template to determine the optimum annealing temperature of **(A)** ORF 9 **(B)** ORF10 **(C)** ORF11 and **(D)** ORF 13 primer sets. A temperature gradient was set up on a PCR machine and identical assays were amplified at different temperatures ranging from 55.2 -74.7 ^{0}C. The molecular size marker used was Hyperladder 1 (Bioline). Expected amplicon size was **(A)**~2345b **(B)**~ 2460b **(C)** ~2483b. **D** shows multiple bands. A band of ~2258b should be seen.

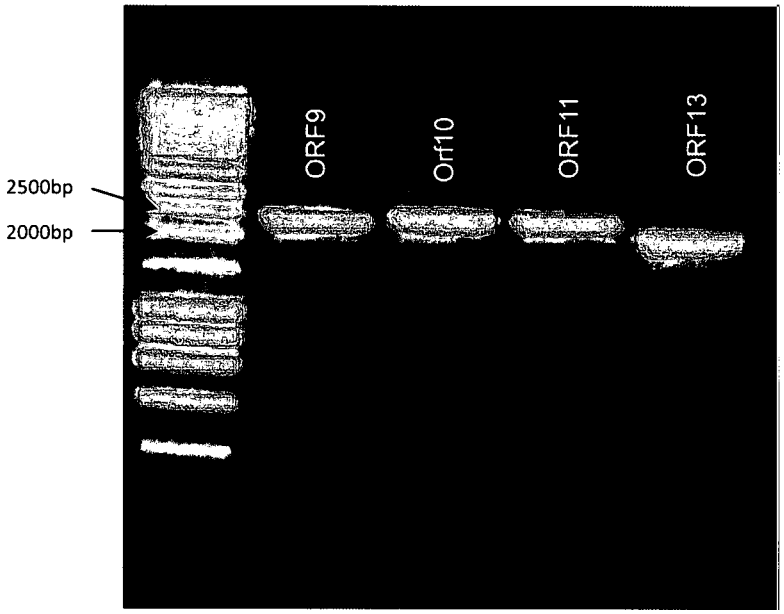


Figure 6.5 PCR amplification products from purified fosmid template DNA

DNA amplified using ORF9, 10, 11 and 13 primer sets at optimised annealing temperatures. The molecular size marker used was Hyperladder 1 (Bioline).

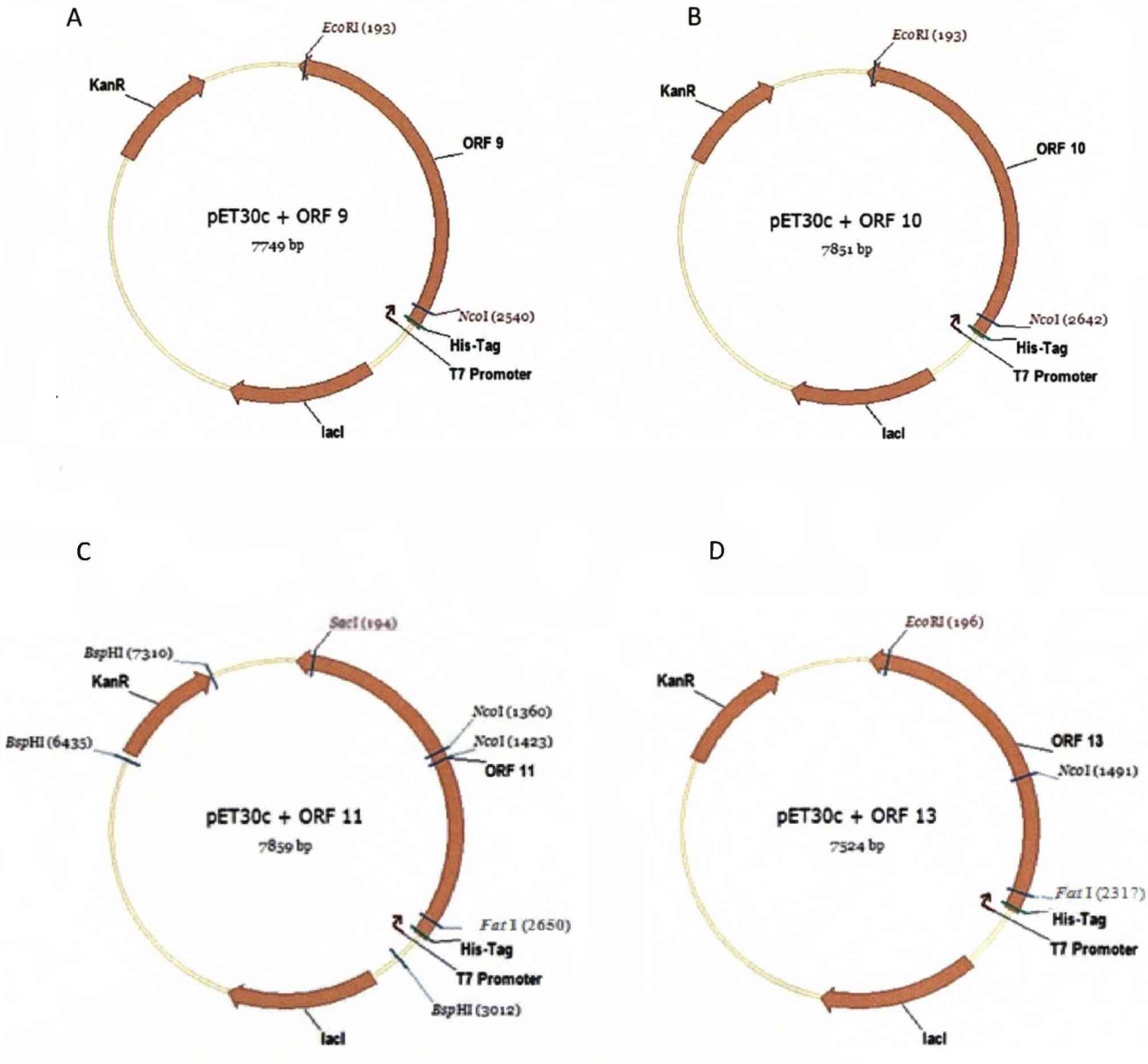


Figure 6.6 pET30c vector constructs for ORF 9, ORF 10, ORF 11 and ORF 13.

A-pET30c expression vector contains ORF 9 with an N-terminal $His_{(6)}$-tag with *Nco* I and *Eco*RI restriction sites and rsistance to kanamycin (*Kan R*).
B-pET30c expression vector contains ORF 10 with an N-terminal $His_{(6)}$-tag with *Nco* I and *Eco*RI restriction sites and resistance to kanamycin (*Kan R*).
C-pET30c expression vector contains ORF 11 with an N-terminal $His_{(6)}$-tag with *FatI* and *Sac*I restriction sites and rsistance to kanamycin (*Kan R*).
D-pET30c expression vector contains ORF 13 with an N-terminal $His_{(6)}$-tag with *Fat*I and *Eco*RI restriction sites and rsistance to kanamycin (*Kan R*).

6.3.7 Screening pET30c + ORF clones for endoglucanase activity

The candidate ORF clones were sub-cloned and screened for endoglucanase activity. The screening method used to screen the fosmid library was applied to screen the individual ORF clones. The following modification was made: ORF clones were inoculated onto a 0.2 µm pore diameter filter (Pall-Gelman), allowing secreted enzymes to pass from the clone cells on to the substrate and nutrients to pass from the media to the clone. Following incubation and sufficient growth of each clone, the colonies could be removed with the membrane and the agar stained without bacterial colonies interfering with the staining process. By this method, only one clone-ORF 10 was found to express endoglucanase activity (Figure 6.7). The presence of endoglucanase activity was checked through zymogram analysis. 10 mL cultures of fosmid clone C4 and ORF 10 were propagated, cells collected and sonicated (6.2.5). 10 µL of each cell extract was separated by 1D SDS-PAGE and gels developed as zymograms (6.2.6). It was confirmed that all the activity expressed by the fosmid clone C4 could be attributed to the ORF 10 gene. Figure 6.7 shows analysis of total cell lysates from fosmid clone C4 and the ORF 10 clone. Although activity of the ORF 10 loaded lane is smeared the 5 distinct bands of activity are clearly visible (Figure 6.7). The five bands from the fosmid clone lysate were accounted for in ORF 10 by zymogram analysis.

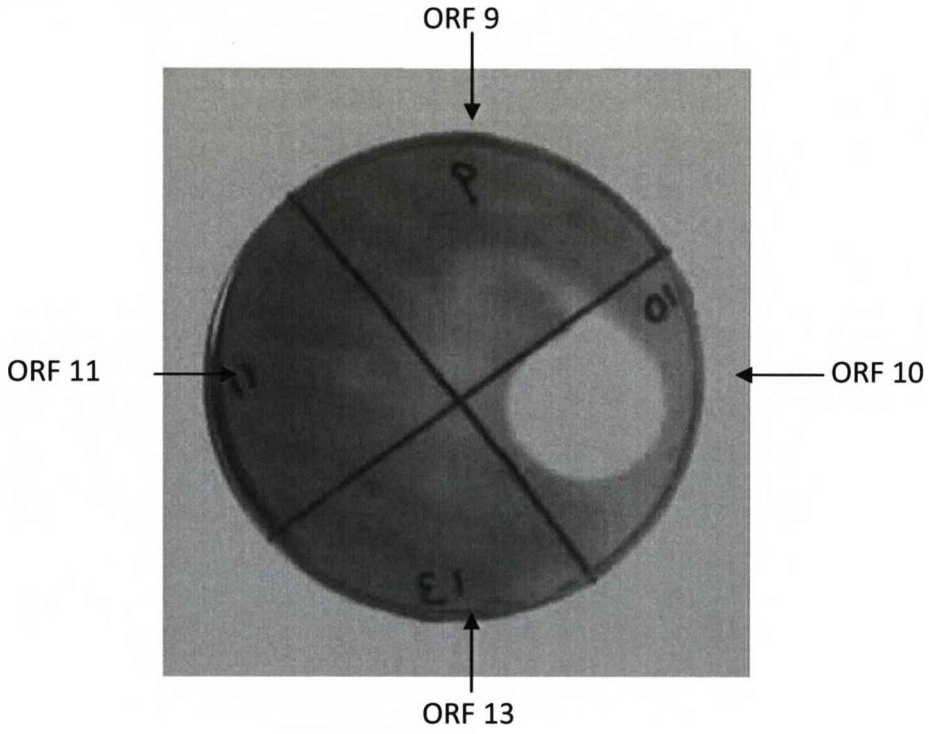


Figure 6.7 Screening of cloned ORFs 9, 10, 11 and 13 for endoglucanse activity

ORF 9, 10, 11 and 13 clones were inoculated onto a 2 µm pore diameter filter (PALL) placed on LB agar (1.8 % w/v), supplemented with CMC (0.1% w/v) and kanamycin (50 µg ml^{-1}). Following incubation the filter was removed and the agar stained with Congo red (0.1 % w/v) for 30 min with shaking. The Congo red was removed and the agar washed with NaCl (1 M), 30 min, changing the NaCl twice. Endoglucanase activity can be seen at clone ORF 10.

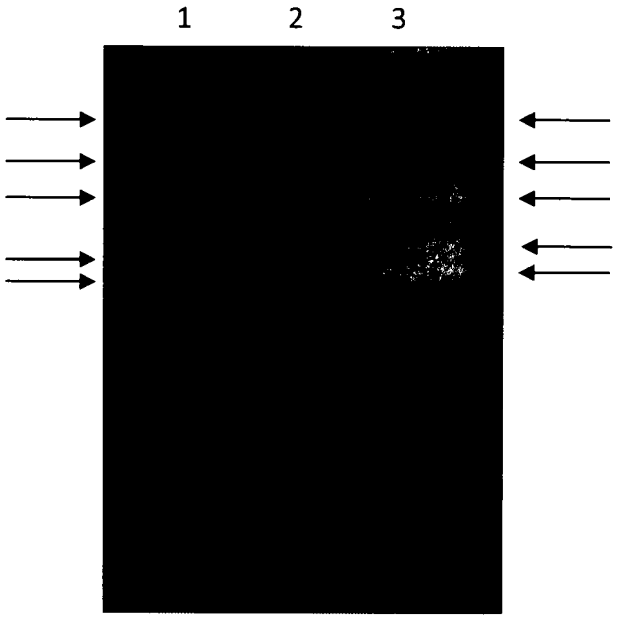


Figure6.7 Zymogram analysis of fosmid clone C4 and pET30c+ORF 10 clone endoglucanase activity

Lane 1: Fosmid clone C4. **Lane 2**: *E.coli* BL21 cells. **Lane 3**: pET30c+ORF 10 clone. Cell lysates prepared by sonication and proteins fractionated on a CMC zymogam. The proteins were washed 2 x 20 min in ddH_2O and incubated in 100 ml 0.1 M sodium phosphate buffer, pH 6.4 with 0.1 % v/v Triton X-100, and incubated for 20 h at 37°C. The gel was stained with Congo red and destained with 1 M NaCl to detect endoglucanase activity. Arrows show protein bands with endoglucanase activity.

6.3.8 Induction and purification of ORF 10 recombinant protein

Clone ORF 10 and *E.coli* strain BL21 (DE3) cells without the cloned insert were grown and induced according to method 6.2.11. 10 µL of each cell lysate was separated by SDS PAGE. Figure 6.8 shows total cell lysates at different time points of induction. A protein of ~100 KDa increased in expression following 1 h induction with IPTG and continued to increase following 2 and 3 h inductions. To act as a control, the induction method was repeated using BL21 (DE3) *E. coli* cells which had been transformed with the pET30c vector lacking the ORF10 insert. The 1D SDS-PAGE shows that the band of ~ 100 kDa is not expressed by the control BL21 (DE3) cells without the presence of the cloned insert.

Purification of the soluble recombinant $His_{(6)}$-ORF 10 protein under non-denaturing conditions was attempted. Although active bands could be visualised using zymogram analysis (Figure 6.7), when the soluble fraction of protein following sonication was analysed using 1D SDS-PAGE no over-expressed proteins were evident. It was thought that inclusion bodies were possibly being formed which is a common tendency with over-expressed recombinant proteins from multi copy plasmids (Simpson 2003). It is thought that inclusion bodies are incorrectly folded or partially folded proteins especially produced by proteins containing disulphide bonds (Simpson, 2003). Solubility of the recombinant protein and its cellular localisation following sonication was examined (6.2.12). When the total protein extract is collected by heating the cell pellet in Laemmli buffer (Lane 1) there is over expression of a ~ 100 kDa protein clearly evident. When the soluble fraction of protein is extracted and then various concentrations of urea was added to the cell debris it was found that a protein of ~100 kDa was relinquished at an increased level from the cell debris as the concentration of urea was increased (Figure 6.9). This provided evidence for the hypothesis that the recombinant protein was present in the insoluble protein fraction. following cell lysis

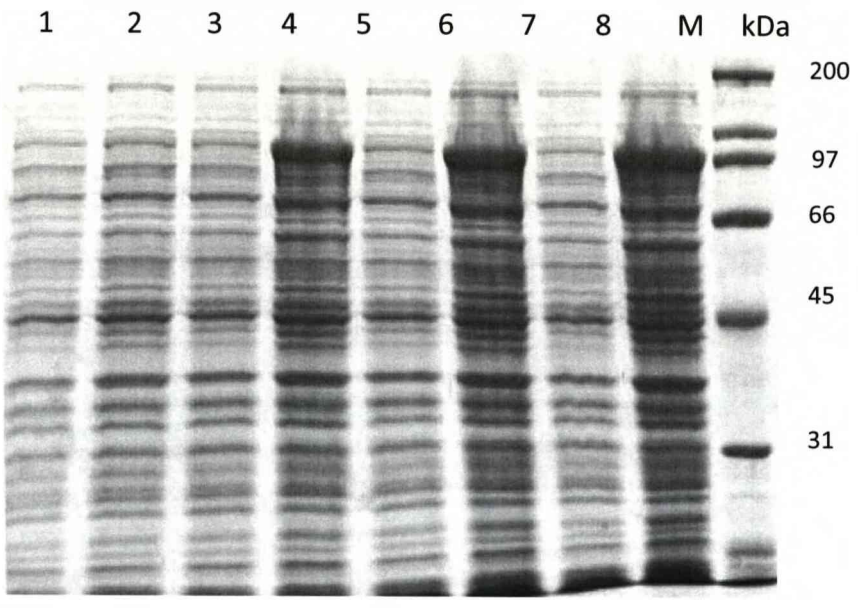


Figure 6.8 SDS-PAGE analysis of total cell lysates of ORF 10 clone and BL21 (DE3) E.coli control

M. Broad range molecular weight marker (BioRad); **Lane 1** BL21 (DE3) *E.coli* lysate at OD_{600} 0.6; **Lane 2** Clone ORF 10 lysate at OD_{600} 0.6; **Lane 3** BL21 (DE3) *E.coli* lysate following 1 h induction; **Lane 4** Clone ORF 10 following 1 h induction; **Lane 5** BL21 (DE3) *E.coli* lysate following 2 h induction; **Lane 6** Clone ORF 10; **Lane 7** BL21 (DE3) *E.coli* lysate following 3 h induction; **Lane 8** Clone ORF 10 following 3 h induction.

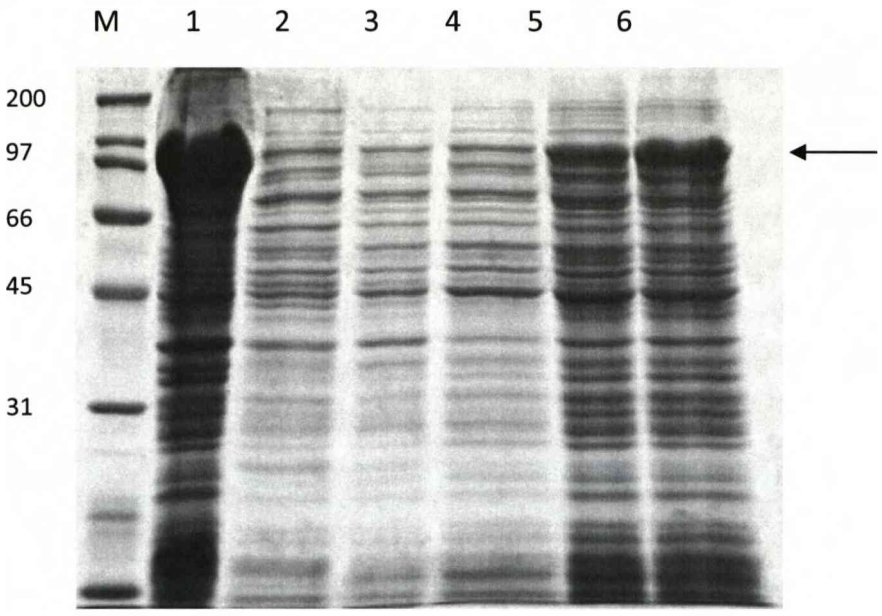


Figure 6.9 SDS-PAGE separation of recombinant $His_{(6)}$-ORF 10 protein to assess solubility of the over expressed protein

M. Broard range molecular weight marker (BioRad); **Lane1** Total protein following heating in1 x Laemmli buffer; **Lane 2** Supernatant of cell lysate following sonication in PBS; **Lane 3** Insoluble portion of lysate in with 2 M urea pellet; **Lane 4** 4 M urea; **Lane 5** 6 M Urea; **Lane 6** 8 mM urea. Arrow indicates presence of a ~100 kDa protein relinquished with increasing urea concentration.

Protein purification of the $His_{(6)}$ - ORF 10 recombinant protein was subsequently carried out according to method 6.2.13, using the His-tag fusion of the recombinant protein for purification based on Ni^{2+} chelation chromatography. Following loading and washing of the pET30c ORF 10 clone protein extract some non-specifically bound *E.coli* proteins eluted with 20 mM imidazole. When fractions were examined by 1D SDS-PAGE a protein of ~ 100 kDa eluted with the 100 mM and 250 mM fractions. Although this was the major protein in these fractions (Figure 6.10), both contained many other proteins. It was thought that the recombinant proteins may not have bound as a result of aggregation. Non-specifically bound proteins were also an issue and in an attempt to circumvent both these problems with purification, DTT (50 mM) was introduced to fully reduce proteins and prevent the formation of disulphide bonded aggregates and in the event that incorrect folding of the recombinant protein made the His-tag inaccessible to binding with the nickel column. Incorrect folding can occur in cysteine-rich proteins. The purification was repeated using 50 mM DTT resulting in a decrease of non-specifically bound proteins. The purest fraction, eluted with 100 mM imidazole was dialysed to remove the imidazole, of which 10 µL was separated by denaturing SDS-PAGE producing a single band of ~100 KDa (Figure 6.11). Denaturing SDS-PAGE was used for the development of a zymogram. Following separation under denaturing conditions the proteins were renatured in the gel and stained for endoglucanase activity using Congo red. A single band of ~100 KDa was produced following 20 h incubation of the gel in renaturation buffer (Figure 6.11).

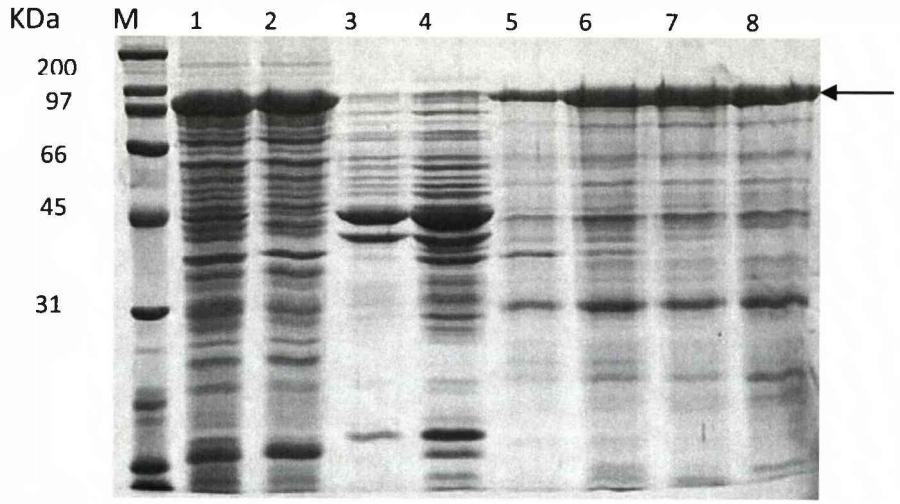


Figure 6.10 SDS-PAGE analysis of His-Trap eluted proteins (without the addition of DTT)

Potein separated by SDS-PAGE. The gel was stained with Coomassie and destained with 10 % (v/v) acetic acid; 10 % (v/v) methanol. **M**, broad range molecular weight marker (BioRad); **1**, Total protein extract; **2**, unbound protein; **3**, Wash 1-unbound protein; **4**, Wash 2-unbound protein; **5**, Elution 1 – 20 mM Imidazole; **6**, Elution 2 – 100 mM Imidazole; **7**, Elution 3 – 100 mM Imidazole; **8**, Elution 4 – 250 mM Imidazole. Arrow indicates overexpressed protein.

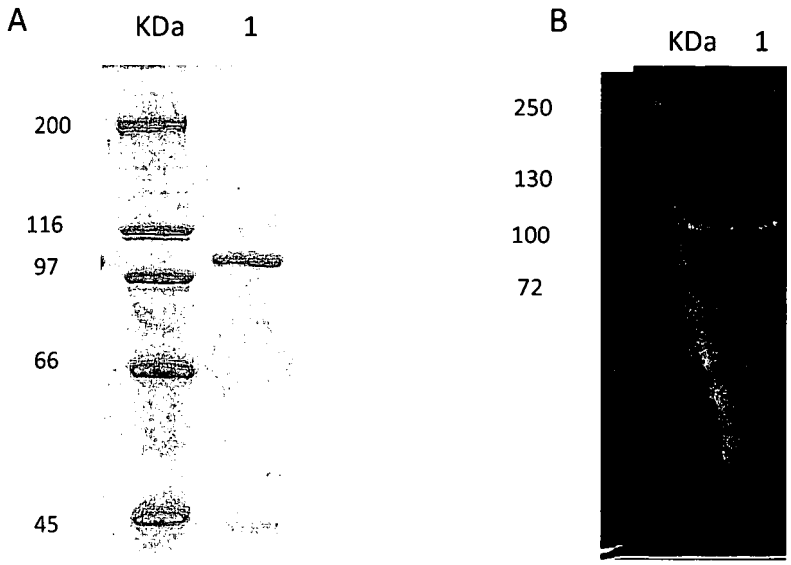


Figure 6.11 His (6)-ORF 10 recombinant purified protein

A fraction eluted with 100 mM imidazole was separated on a 7% SDS-PAGE separating and 4 % stacking polyacrylamide gel.

A, stained with Coomassie and de-stained with 10% methanol; 10% acetic acid. A BioRad broard range molecular weight marker was used.

B, gel developed for zymogram analysis, 0.1% CMC was incorporated into a 7% PAG. The Page Ruler™ Plus prestained protein ladder was used.

6.3.9 MALDI-TOF analysis of the purified $His_{(6)}$ - ORF 10 recombinant protein

The protein band of molecular size ~100 KDa (figure 6.11) was excised from the SDS-PAGE Coomassie stained gel. Following tryptic digestion (see chapter 4 (4.2.5.2)), MALDI-TOF analysis of the tryptic peptides was carried out. The $M + H^+$ value of the peptides detected were used to search against the translated predicted peptide masses for ORF10-His recombinant protein. Peptide masses were calculated by hypothetical digestion with trypsin using the ExPASy PEPTIDEMASS tool (http://www.expasy.ch/tools/peptide-mass-ref.html). This resulted in a number of matching peptide masses (Figure 6.12). Peptides were matched mainly to the N-terminal end of the protein. Analysis of the hypothetically digested sequence showed that the C-terminal end of the protein was largely devoid of tryptic digestion sites (K and R residues). The source of the protein was also from an in-gel tryptic digestion which can yield less peptides than that of an in-solution digestion.

The predicted ORF 10 translated sequence band size is slightly larger than the predicted 92 kDa His-tagged protein taking in to account the translational start and stop codons of the vector calculated (Figure 6.6) using the ExPASy proteomics server molecular weight calculator.

```
        10          20          30          40          50          60
MHHHHHHSSG LVPRGMKETA AAKFERQHMD SPDLGTDDDD KAMKTEQEFI KKGLVLGVSL
        70          80          90         100         110         120
LLMAIIIMSS TPKTSANSAN PYLWPYNQST NISFNESDVY DAWTAWRDAQ ITSNNAGGNG
       130         140         150         160         170         180
RYRVMGGVDG GSTVSEGQAY GILYTSIFDE QTLFDGLFLF AKDHYNTQGV MDWHIGSPGV
       190         200         210         220         230         240
RIGSGGATDA EVDMAMGLVN ACVKVQQNAW SASSAGIDYC VEATNLINAI YTYEVDHAGS
       250         260         270         280         290         300
SPPGGLPNNQ GNELLPGDTW DVSGTYPDGI INLSYFPPGY FTVFGKFTQN EAAWNAVIDR
       310         320         330         340         350         360
NYEVTDLVQA KSDNCSGLVP NWNKYNGDAQ LVSWQTNNYS WWSYDAARFA WRIAVDQAWY
       370         380         390         400         410         420
GRPEATETMN EIGGFFSSTG FNNIGEHSMN GIKTGSGPWP FFVANAASAV WAAPNPVATN
       430         440         450         460         470         480
CGTGTGSLQE SPQSAYNRVL STKDNPNSYY GNAWRLFSML LMTGNFPNFY EMADGNVTPV
       490         500         510         520         530         540
PTSTPGSATN TPVPPTATIP AGTGACHVDY VVANEWGSGF QANVTITNNM SSAIDGYTLT
       550         560         570         580         590         600
WTHAPGQVVS SGWNVTVSQT GNQVTATNPA GSWMGKINAN GGTSSFGFQG SLTSKAVVPT
       610         620         630         640         650         660
DFVLNGTACN GDTPPTETPI PTPETPTVTP TVWCPQATSV PLVVEPVTSP TNELSQTLVV
       670         680         690         700         710         720
KVYADWVSAT GPAGSVTVDT PEADGFHVTV PLAANSINNI SVKSQIPVVT NPNGCTYGGY
       730         740         750         760         770         780
TLSKTVTIVQ ESDAVTPTLT PTATATATAT PTATATTPSG TATCSVAYTV GNDWGSGFTT
       790         800         810         820         830         840
DVKITNKGAS TINGWTLTYT YAGNQTITNA WNATVTQSGK TITATDAGWN GTLPPNGSAS
       850         860
FGFQGSYSGS NIAPTTFKVN GSVCQ
```

Obtained mass	Missed Cleavage	Predicted mass [M+H$^+$]	Peptide position
3897.6	4	3897.8	18-51
2285.5	1	2284.4	52-73
3967.4	0	3968.8	74-107
1635.4	0	1634.8	287-300
2897.0	0	2895.3	325-348
1457.8	0	1456.6	444-455

Figure 6.12 MALDI-TOF analysis of $His_{(6)}$ – ORF 10 recombinant protein.
A protein band of ~100 kDa believed to be the purified $His_{(6)}$ – ORF 10 recombinant protein was excised from a PA gel, digested with trypsin and subjected to MALDI-TOF analysis. M+H$^+$ values were determined with and without missed cleavages for a number of peptides (obtained column) and matched against a hypothetic tryptic digest of the $His_{(6)}$ – ORF 10 recombinant protein (Predicted mass [M+H$^+$]). The residue position of each matched peptide is indicated by underlining of the protein sequence and peptide position column.

6.4 Discussion

As there is a limited number of bacterial species which can be successfully cultivated under standard laboratory conditions, large insert environmental DNA libraries provide an advantageous molecular tool for isolation and characterisation of enzymes that may be active in the environment but unknown in collections of laboratory cultures. Fosmids can offer stable low copy number gene expression which can be greatly increased upon induction, limiting the toxic effects that products may have on a heterologous host and increasing the chance of successful cloning and identification of genes encoding novel enzymes. The method developed offers an attractive prospect for screening large insert libraries for truly novel enzyme encoding genes, as sequence information is unnecessary. However this approach does require that the gene is transcribed, translated and folded correctly In the host and while this may appear to be a major drawback, the results presented here demonstrate that success can be achieved.

Metagenomic libraries have been constructed and screened for cellulase activity previously. A cosmid library constructed from soil DNA enriched for agarolytic activity yielded 8 endoglucanse positives from 1700 clones screened (Voget *et al.*, 2006) using a Congo red / agar plate method similar to that described here. Another cosmid library was constructed with DNA from compost yielded 4 positive endoglucanse clones from ca. 100,000 clones screened. A Congo red / agar plate method was also used for screening (Pang *et al.*, 2009). A further metagenomic library of 70,000 clones with inserts of ~ 40 kb was constructed from soil DNA yielding only one endoglucanse positive clone, identified also using a similar Congo red / agar plate method (Kim *et al.*, 2008).

In this experiment a fosmid library of ca. 7000 clones was constructed using DNA extracted from estuarine sediment organisms and from which one positive clone was identified.

Previous fosmid library-based, metagenome studies have located phylogenetic markers such as 16 S rRNA genes (Woebken *et al.*, 2007; Gilbert *et al.*, 2008), this was not the case with fosmid clone C4. There is however the opportunity to design primers for

regions of the sequenced fosmid and probe the library for fosmids with overlapping sequences, which may elucidate the phylogenetic origins of the organism from which the functional genes have originated. Alternatively, or additionally a general taxonomic classification could be elucidated based on the homology of the fosmid sequence to those known in the reference databases. As the architecture and sequence of ORF 10 is novel in comparison to reference sequences it would possibly be advantageous to identify which species the fragment of DNA has originated. Identifying DNA fragments from the same native host either by isolation or screening the library further using probes designed for the fosmid clone DNA fragment may elucidate a full polysaccharide-degrading system with exploitable features. Unfortunately further work on this study was restricted by time.

Sequence analysis of ORF 10 showed that the closest match available in the nr protein database was one present in *Herpatosiphon aurantiacus*, although ORF 10 contains 2 CBMs whereas that of *H. aurantiacus* contains only one. The two proteins share 54 % sequence identity. *H. aurantiacus* is a filamentous gliding bacterium of the class *Flexibacterales* and has been isolated from a number of habitats including marine shores (Holt & Lewin, 1968). Although the purified $His_{(6)}$ - ORF 10 recombinant protein was obtained, the need to purify the protein under denaturing conditions limited the biochemical characterisation of the protein. For example, information on temperature, pH relationships, substrate specificities and range could not be determined. It could be suggested that the clone should be transformed into another host system, as the host strain could be forming inclusion bodies possibly resulting from the evolutionary distance between the ORF 10 gene and the host strain, making it difficult to obtain the soluble recombinant protein for further biochemical investigation. Following analysis of the ORF 10 sequence, it was observed that a number of cysteine residues (10) are present. Cysteine rich proteins may result in incorrect folding and S-S bond formation, causing aggregation of the protein (Simpson, 2003). This was probably the case with the recombinant $His_{(6)}$-ORF 10 protein as the denaturation and unfolding of the protein using the reducing agent DTT significantly increased the yield of purified recombinant protein. It

has previously been shown that the positioning of His-tags in recombinant proteins effect protein solubility (Xu *et al.*, 2008) and the production of an ORF 10-His $_{(6)}$ instead of a recombinant His $_{(6)}$-ORF 10 protein may provide better protein folding making the His $_{(6)}$ site available for binding to the Ni column.

The molecular weight of the translated ORF 10 protein sequence as predicted by the Artemis program was 87 kDa. The predicted molecular weight of the recombinant His $_{(6)}$ – ORF 10 protein as calculated using the MW/pI compute program at ExPASy was 92 kDa. The molecular weight of the actual expressed purified recombinant His $_{(6)}$ – ORF 10 protein observed when fractionated by SDS-PAGE and zymogram analysis was ~100 kDa. The slightly larger MW of the protein may be also explained by post-translational modification by the host such as glycosylation. Although SDS-PAGE provides molecular weight estimations of proteins within 10 % of the actual molecular weight (Simpson, 2003).

Unfortunately there was insufficient time to follow up the predicted activities of ORF9, ORF 11 and ORF13. The ORF 9 and 11 sequences contained predicted conserved domains of GH family 6 and 5 respectively. Other substrates could have been used to screen for activity such as Xylan or different length oligosaccharides. Family 6 glycosyl hydrolases contain enzymes with known β-1, 4-endoglucanase and exoglucanase activity. Whilst family 5 GHs have a wide range of activities involved in the degradation of the cellulose polymer (http://www.expasy.org/cgi-bin/lists?glycosid.txt; http://www.cazy.org; Cantarel *et al.*, 2009). Although activity against CMC was not observed, it would be predicted that both these genes are involved in depolymerisation of cellulose at some stage. This might be further reason for identifying the organism from which the fosmid sequence originated.

The ORF 13 sequence interestingly contains in addition, to a peptidase family domain, two CBM family 3 domains. This family of CBMs are known to bind to cellulose as constituents of endoglucanases (http://www.cazy.org). It is hypothesised that the peptidase catalytic domain may be involved in the breakdown of proteinaceous material

associated with cellulose fibres in the native host. This has been reported for chitinolytic activity of *Alteromonas* sp. Strain 0-7, where several chitinases are expressed with a protease containing a chitin-binding module following chitin induction and is suggested to be an essential part of complete degradation of crustacean cuticles where chitin is associated with proteins, lipids and calcium carbonate (Miyamoto *et al.*, 2002). Alternatively, the native host may express the peptidase to modulate the binding and catalysis of the glycosyl hydrolases.

6.5 Conclusions

- A method for screening endoglucanase activity encoded on a large insert fosmid library was developed.
- A library of 7104 Clones of DNA extracted from Colne estuary sediment was screened resulting in the identification of one clone that expressed endoglucanase activity.
- The fosmid was sequenced, assembled and annotated.
- Primer sets for four predicted ORFs were designed and each ORF cloned and transformed into *E.coli* .
- Each sub-clone was screened for endoglucanase activity and one, ORF10, was shown to be positive.
- The recombinant ORF10-His protein was over expressed and purified to homogeneity using a His-trap column under denaturing conditions.

Chapter 7

General Discussion

Approximately 71 % of the Earth's surface is covered by ocean (Suttle, 2007), where half the global primary production occurs, via the action of phototrophic bacteria and algae which harvest solar energy to produce organic matter fuelling heterotrophic processes in the ocean (Karl, 2007). Polysaccharides, proteins and lipids produced as a result of primary production as well as senescence, moulting or excretion from marine organisms form part of the particulate organic matter (POM) and dissolved organic matter (DOM) pools in the marine environment. Cellulose and chitin are collectively the two most abundant polysaccharides on Earth and degradation of these recalcitrant substrates is therefore an important driving process of the carbon cycle. Although many cultured bacterial species are known to degrade cellulose and chitin, little is known of the uncultured fraction particularly in the marine environment where there is a clear paucity of information on true community function and structure (Fuhrman & Hagstrom, 2008). What is known of bacterial species in laboratory culture does not necessarily equate with the functions of bacterial species when part of a microbial community in the environment.

The glycosyl hydrolases (GH's) are a diverse and abundant group of enzymes, produced by Bacteria, Eukarya and Archaea to hydrolyse the β 1,4- glycosidic bond of complex polysaccharides to release soluble oligo- and mono-saccharides. Thus far there are 115 representative families of glycosyl hydrolases, most of which originate from cultured microbial species (Cantarel *et al.*, 2009).

Recent developments in molecular techniques have enabled a previously unattainable insight into the abundance and diversity of key members of microbial communities in the marine environment. Here metagenomics was used to assess the bacterial community colonising cellulose bait in the Irish Sea, allowing for the first time the analysis of key species involved, the community structure and the repertoire of GH

genes that are present in that community. A method previously used to isolate DNA from cellulose baits (McDonald *et al.*, 2009) was used to extract DNA from the community biofilm that had colonised the cellulose bait from the Irish Sea. Following random 454 pyrosequencing, reads were assembled into 26,860 contigs which were taxonomically and functionally assigned using the metagenomic facilities MEGAN (Huson *et al.*, 2007) and MG-RAST (Meyer *et al.*, 2008). Both sets of analysis, based on similarity matches using blastx searches against two different reference databases confirmed that Bacteria were the dominant group (69 % and 61 % of genes respectively) in the biofilm community. Numerically dominant phyla represented by contigs assigned to the Bacteria were by far the Proteobacteria (55 % and 63 %) and the *Bacteroidetes/ Chlorobium* group (27 % and 31 %). Within these phyla large proportions of genes were assigned to species with well characterised cellulase systems-*Saccharophagus degradans, Cellvibrio japonicus* and *Cytophaga hutchinsonii.* However, as two of these organisms were previously isolated from soil (Deboy *et al.*, 2008; Xie *et al.*, 2007) it is apparent that only general trends regarding taxonomic and metabolic diversity, and abundance can be concluded in the complex bacterial biofilm.

Analysis of the dataset by comparison to 16S rRNA databases provides an alternative to taxonomic assignment based on similarity to functional genes. 16S rRNA genes are used as molecular markers as they are present in all prokaryotes and the ribosomal subunits shows functional consistency and therefore high levels of sequence conservation. Evidence for horizontal transfer of rRNA genes is also limited (Röling & Head, 2005). Comparison to the Greengenes database revealed fourteen matches of gene fragments to sequences in the database, further support the evidence that the community was dominated by the *Proteobacteria* and *Bacteroidetes*, with ten matches to the *Proteobacteria* and four to the *Bacteroidetes* providing approximately the same ratio of 2:1 for hits to *Proteobacteria* and *Bacteroidetes* as the functional gene based assignments. Interestingly, three fragments were matched against the Greengenes database and two against the RDP database for *Glaciecola spp.* This genus was included in taxonomy

following the isolation of two phychrophilic Gram negative bacterial isolates from sea-ice diatom assemblages at coastal regions of eastern Antarctica (Bowman *et al.*, 1998). Although cellulose hydrolysis has not been reported in this group, xylanase and agarase production has been documented for members of the genus (Romanenko *et al.*, 2003; Yong *et al.*, 2007; Guo *et al.*, 2009). Further evidence to substantiate the claim that *Glaciecola* spp are important members of the cellulose degrading community was provided by the isolation of a small collection of strains of this genus from the biofilm colonising cellulose bait in the Irish Sea. Screening of these isolates using carboxymethyl cellulose (CMC), showed the expression of endoglucanase from six phylogenetically inferred *Glaciecola* strains. Future work using qPCR could decipher the relative abundance of these endoglucanase-producing *Glaciecola* strains in the cellulose colonising biofilm.

Further analysis of the genetic potential of GHs produced by the community provided a number of hits to *Saccharophagus degradans, Cellvibrio japonicus* and *Cytophaga hutchinsonii.* Again, this is most likely because these are species with fully annotated genomes and well-characterised cellulase systems rather than the species are actually present. A total of 116 of the contigs provided matches to the constructed GH database. Further work on a comprehensive GH database including all exo-acting cellulases and β-glycosidases to examine the full polysaccharide degrading gene repertoire of the community is required. Additionally, sequences closely related to cohesins and dockerins (proteins associated with cellulosome complexes) have recently been identified in the *Saccharophagus degradans* genome (Weiner *et al.*, 2008). Therefore, these together with carbohydrate binding modules could also have been investigated. Metagenomics is clearly a valuable tool for providing an initial insight into the population dynamics of environmental samples. However there are obvious drawbacks obvious in assembling short reads and matching to a limited reference database. It is hoped that advances in technology will circumvent these problems with longer pyrosequencing reads and a more comprehensive coverage of annotated genomes, especially with recent progress in single cell genomics (Zhang *et al.*, 2006; Woyke *et al.*, 2009). More useful

information may yet be retrieved from communities by random sequencing of RNA for metatranscriptomics to decipher both phylogenetic and functional diversity within the community. Demonstrating gene expression in the biofilm is the next step and could provide a means to determine community responses to environmental stresses such as pollution and climate change.

With the growth in genomic data becoming available for microbial communities, metaproteomics has the potential to provide additional functional information. The technique has proven successful when applied to communities with well characterised metagenomes and limited microbial diversity (Ram *et al.*, 2005; Lo *et al.*, 2007). However whilst some communities are proving ideal for metaproteomic characterisation with protein identification possible at the strain level (Lo *et al.*, 2007; Wilmes *et al.*, 2008) only inferences as to the general functions of the colonised cellulose community could be made in this project, with a large proportion of peptides matching proteins involved in carbohydrate metabolism. Initial difficulties in protein extraction and purification were compounded by the time involved in chromatographic separation, MS/MS and manual *de novo* peptide sequencing. New ion-trap MS coupled with algorithms for mass spectra searches, abolishing the need for *de novo* sequencing when comprehensive metagenome data are available, should ultimately provide an excellent platform development in this field e.g. Lo *et al.* (2007).

A fosmid library of ca. 7000 clones was obtained from Dr Ashley Houlden (University of Sheffield) and functionally screened for endoglucanase activity. The aim was to develop a functional screening method for a large insert clone library constructed from marine environmental DNA. One clone expressing endoglucanse activity was sequenced and annotated, enabling the PCR amplification and subcloning of four predicted ORFs. Endoglucanse activity was located to one ORF with sequence similarity to one GH8 family domain and two CBM 2 family domains. This approach has potential for locating enzymes involved in key functions, not only for improved understanding of the community and its activity but also for the exploitation of enzymes in the biotechnology industry. There are

advantages of large insert clone libraries over the random pyrosequencing of environmental DNA, in that characterisation of expressed proteins from predicted genes can be achieved rather than simply using *in silico* similarity based identification. The latter can often be misleading and limited to the context of reference databases. However random pyrosequencing based methods do have the advantage of producing an otherwise unobtainable plethora of information on community structure. Also, extraction of intact DNA from environmental communities is often difficult and large fragments of DNA are not required for pyrosequencing. The results provide a fresh insight into the specialised bacterial community colonising the insoluble polysaccharide cellulose, whose hydrolysis is central to the carbon cycle.

The importance of *Glaciecola* spp has previously been described to some extent with members of this genus previously found in association with diatom species (Bowman *et al.*, 1998), an important source of cellulose and other polysaccharides. However their role in cellulose degradation has not previously been described. The number of sequence matches of cellulose biofilm DNA against *Glaciecola* rRNA sequences suggests that they are a significant genus in this community, and isolates collected from the same source were shown here to express endoglucanase, further supporting their involvement in the attached degrading consortium.

Finally, for the first time a marine biofilm colonising polysaccharide has been subjected to characterisation of *in situ* expressed proteins. Whilst the species origins of the proteins could not be determined, cellulases and chitinases are present and carbohydrate metabolism is clearly a dominant function as evidenced by metaproteomic data. An additional insight in to the functioning of the bacterial community colonising the cellulose baits was provided by SEM of the bait itself. Clear signs of degradation of the polysaccharide can be seen by an array of largely rod shaped bacteria.

Taken together, therefore, the importance of Bacteria as primary colonisers of insoluble polysaccharides in the marine environment has been directly demonstrated.

Chapter 8

References

Aebersold, R., and M. Mann. (2003) Mass spectrometry-based proteomics. *Nature* **422:** 198-207.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. (1990) "Basic local alignment search tool." *J. Mol. Biol.* **215**: 403-410

Alzari, P. M., H. Souchon, and R. Dominguez. (1996) The crystal structure of endoglucanase CelA, a family 8 glycosyl hydrolase from Clostridium thermocellum. *Structure* **4:** 265-75.

Amann, R. I., W. Ludwig, and K. H. Schleifer. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59:** 143-69.

Andronopoulou, E., and C. E. Vorgias. (2003) Purification and characterization of a new hyperthermostable, allosamidin-insensitive and denaturation-resistant chitinase from the hyperthermophilic archaeon *Thermococcus chitonophagus*. *Extremophiles* **7:** 43-53.

Aronson, J. M., and M. S. Fuller. (1969) Cell wall structure of the marine fungus, *Atkinsiella dubia*. *Arch Microbiol* **68:** 295-305.

Ashelford, K. E., N. A. Chuzhanova, J. C. Fry, A. J. Jones, and A. J. Weightman. (2005) At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. *Appl Environ Microbiol* **71:** 7724-36.

Azam, F., and R. A. Long. (2001) Sea snow microcosms. *Nature* **414:** 495, 497-8.

Azam, F., and F. Malfatti. (2007) Microbial structuring of marine ecosystems. N*at Rev Microbiol* **5:** 82-91.

Aziz, R. K., D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, K. Formsma, S. Gerdes, E. M. Glass, M. Kubal, F. Meyer, G. J. Olsen, R. Olson, A. L. Osterman, R. A. Overbeek, L. K. McNeil, D. Paarmann, T. Paczian, B. Parrello, G. D. Pusch, C. Reich, R. Stevens, O. Vassieva, V. Vonstein, A. Wilke, and O. Zagnitko. (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9:** 75.

Baik, K. S., Y. D. Park, C. N. Seong, E. M. Kim, K. S. Bae, and J. Chun. (2006) *Glaciecola nitratireducens* sp. nov., isolated from seawater. *Int J Syst Evol Microbiol* **56:** 2185-8.

Banfield, J. F., N. C. Verberkmoes, R. L. Hettich, and M. P. Thelen. (2005) Proteogenomic approaches for the molecular characterization of natural microbial communities. *OMICS* **9:** 301-33.

Bauer, M., M. Kube, H. Teeling, M. Richter, T. Lombardot, E. Allers, C. A. Wurdemann, C. Quast, H. Kuhl, F. Knaust, D. Woebken, K. Bischof, M. Mussmann, J. V. Choudhuri, F. Meyer, R. Reinhardt, R. I. Amann, and F. O. Glockner. (2006) Whole genome analysis of the marine *Bacteroidetes Gramella forsetii* reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol* **8:** 2201-13.

Bayer, E. A., R. Lamed, B. A. White, and H. J. Flint. (2008) From cellulosomes to cellulosomics. *Chem Rec* **8:** 364-77.

Bayer, E. A., E. Setter, and R. Lamed. (1985) Organization and distribution of the cellulosome in *Clostridium thermocellum*. *J Bacteriol* **163:** 552-9.

Bayer, E. A., L. J. Shimon, Y. Shoham, and R. Lamed. (1998) Cellulosomes-structure and ultrastructure. *J Struct Biol* **124:** 221-34.

Benndorf, D., G. U. Balcke, H. Harms, and M. von Bergen. (2007) Functional metaproteome analysis of protein extracts from contaminated soil and groundwater. *ISME J* **1:** 224-34.

Bhattacharya, D., A. Nagpure, and R. K. Gupta. (2007) Bacterial chitinases: properties and potential. *Crit Rev Biotechnol* **27:** 21-8.

Biddle, J. F., C. House, S. Fitz-Gibbon, S. Schuster, and J. Brenchley. (2007) Metagenomics of deeply buried marine sediments. *Geochimica Et Cosmochimica Acta* **71:** A90-A90.

Blackwell, J., K. D. Parker, and K. M. Rudall. (1967) Chitin fibres of the diatoms *Thalassiosira fluviatilis* and *Cyclotella cryptica*. *J Mol Biol* **28:**383-5.

Bolam, D. N., A. Ciruela, S. McQueen-Mason, P. Simpson, M. P. Williamson, J. E. Rixon, A. Boraston, G. P. Hazlewood, and H. J. Gilbert. (1998) Pseudomonas cellulose-binding domains mediate their effects by increasing enzyme substrate proximity. *Biochem J* **331:** 775-81.

Boraston, A. B., D. N. Bolam, H. J. Gilbert, and G. J. Davies. (2004) Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemical Journal* **382:** 769-781.

Bowman, J. P., S. A. McCammon, J. L. Brown, and T. A. McMeekin. (1998) *Glaciecola punicea* gen. nov., sp. nov. and *Glaciecola pallidula* gen. nov., sp. nov.: psychrophilic bacteria from Antarctic sea-ice habitats. *International Journal of Systematic Bacteriology* **48:** 1213-1222.

Bowman, J. P., S. A. McCammon, T. Lewis, J. H. Skerratt, J. L. Brown, D. S. Nichols, and T. A. McMeekin. (1998) *Psychroflexus torquis* gen. nov., sp. nov., a psychrophilic species from Antarctic sea ice, and reclassification of *Flavobacterium gondwanense* as *Psychroflexus gondwanense* gen. nov., comb. nov. *Microbiology* **144:** 1601-9.

Boyer, J. N. (1994) Aerobic and anaerobic degradation and mineralization of 14C-chitin by water column and sediment inocula of the York River estuary, Virginia. *Appl Environ Microbiol* **60:** 174-9.

Bravo-Linares, C. M., S. M. Mudge, and R. H. Loyola-Sepulveda. (2007) Occurrence of volatile organic compounds (VOCs) in Liverpool Bay, Irish Sea. *Mar Pollut Bull* **54:** 1742-53.

Brulc, J. M., D. A. Antonopoulos, M. E. Miller, M. K. Wilson, A. C. Yannarell, E. A. Dinsdale, R. E. Edwards, E. D. Frank, J. B. Emerson, P. Wacklin, P. M. Coutinho, B. Henrissat, K. E. Nelson, and B. A. White. (2009) Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proc Natl Acad Sci* **106:** 1948-53.

Buchan, A., J. M. Gonzalez, and M. A. Moran. (2005) Overview of the marine *Roseobacter* lineage. *Appl Environ Microbiol* **71:** 5665-77.

Button, D. K., F. Schut, P. Quang, R. Martin, and B. R. Robertson. (1993) Viability and Isolation of Marine Bacteria by Dilution Culture: Theory, Procedures, and Initial Results. *Appl Environ Microbiol* **59:** 881-891.

Cantarel, B. L., P. M. Coutinho, C. Rancurel, T. Bernard, V. Lombard, and B. Henrissat. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* **37:** 233-8.

Cauchie, H. M., M. F. Jaspar-Versali, L. Hoffmann, and J. P. Thome. (2002) Potential of using Daphnia magna (crustacea) developing in an aerated waste stabilisation pond as a commercial source of chitin. *Aquaculture* **205:** 103-117.

Celis, J. E., and P. Gromov. (1999) 2D protein electrophoresis: can it be perfected? *Curr Opin Biotechnol* **10:** 16-21.

Chen, L. P., H. Y. Xu, S. Z. Fu, H. X. Fan, Y. H. Liu, S. J. Liu, and Z. P. Liu. (2009) *Glaciecola lipolytica* sp. nov., isolated from seawater near Tianjin city, China. *Int J Syst Evol Microbiol* **59:** 73-6.

Cho, B. C., and F. Azam. (1988) Major Role of Bacteria in Biogeochemical Fluxes in the Oceans Interior. *Nature* **332:** 441-443.

Cohen-Kupiec, R., and I. Chet. (1998) The molecular biology of chitin digestion. *Curr Opin Biotechnol* **9:** 270-7.

Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje. (2005) The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Res* **33:** 294-6.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai, R. J. Farris, A. S. Kulam-Syed-Mohideen, D. M. McGarrell, T. Marsh, G. M. Garrity, and J. M. Tiedje. (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** 141-5.

Collins, J., and B. Hohn. (1978) Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. *Proc Natl Acad Sci* **75:** 4242-6.

Connon, S. A., and S. J. Giovannoni. (2002) High-throughput methods for culturing microorganisms in very-low-nutrient media yield diverse new marine isolates. *Appl Environ Microbiol* **68:** 3878-85.

Cottrell, M. T., and D. L. Kirchman. (2000) Natural assemblages of marine proteobacteria and members of the Cytophaga-Flavobacter cluster consuming low- and high-molecular-weight dissolved organic matter. *Appl Environ Microbiol* **66:** 1692-7.

Cottrell, M. T., D. N. Wood, L. Yu, and D. L. Kirchman. (2000) Selected chitinase genes in cultured and uncultured marine bacteria in the alpha- and gamma-subclasses of the proteobacteria. *Appl Environ Microbiol* **66:** 1195-201.

Cottrell, M. T., L. Yu, and D. L. Kirchman. (2005) Sequence and expression analyses of *Cytophaga*-like hydrolases in a Western arctic metagenomic library and the Sargasso Sea. *Appl Environ Microbiol* **71:** 8506-13.

d'Enfert, C., A. Ryter, and A. P. Pugsley. (1987) Cloning and expression in *Escherichia coli* of the Klebsiella pneumoniae genes for production, surface localization and secretion of the lipoprotein pullulanase. *EMBO J* **6:** 3531-8.

Dabrowski, T., and M. Hartnett. (2008) Modelling travel and residence times in the eastern Irish Sea. *Mar Pollut Bull* **57:** 41-6.

Dalisay, D. S., J. S. Webb, A. Scheffel, C. Svenson, S. James, C. Holmstrom, S. Egan, and S. Kjelleberg. (2006) A mannose-sensitive haemagglutinin (MSHA)-like pilus promotes attachment of *Pseudoalteromonas tunicata* cells to the surface of the green alga *Ulva australis*. *Microbiology* **152:** 2875-83.

Davies, G., and B. Henrissat. (1995) Structures and mechanisms of glycosyl hydrolases. *Structure* **3:** 853-9.

Davies, G. J., K. S. Wilson, and B. Henrissat. (1997) Nomenclature for sugar-binding subsites in glycosyl hydrolases. *Biochem J* **321:** 557-9.

Dean, F. B., S. Hosono, L. Fang, X. Wu, A. F. Faruqi, P. Bray-Ward, Z. Sun, Q. Zong, Y. Du, J. Du, M. Driscoll, W. Song, S. F. Kingsmore, M. Egholm, and R. S. Lasken. (2002) Comprehensive human genome amplification using multiple displacement amplification. *Proc Natl Acad Sci* **99:** 5261-6.

DeBoy, R. T., E. F. Mongodin, D. E. Fouts, L. E. Tailford, H. Khouri, J. B. Emerson, Y. Mohamoud, K. Watkins, B. Henrissat, H. J. Gilbert, and K. E. Nelson. (2008) Insights into plant cell wall degradation from the genome sequence of the soil bacterium *Cellvibrio japonicus*. *J Bacteriol* **190:** 5455-63.

Delahunty, C., and J. R. Yates, 3rd. (2005) Protein identification using 2D-LC-MS/MS. *Methods* **35:** 248-55.

DeLong, E. F. (2007) Modern microbial seascapes. *Nat Rev Microbiol* **5:** 755-7.

Delong, E. F., D. G. Franks, and A. L. Alldredge. (1993) Phylogenetic Diversity of Aggregate-Attached Vs Free-Living Marine Bacterial Assemblages. *Limnology and Oceanography* **38:** 924-934.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie, K. Keller, T. Huber, D. Dalevi, P. Hu, and G. L. Andersen. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology* **72:** 5069-5072.

Desvaux, M. (2005) *Clostridium cellulolyticum*: model organism of mesophilic cellulolytic *Clostridia*. *FEMS Microbiol Rev* **29:** 741-64.

Doi, R. H., A. Kosugi, K. Murashima, Y. Tamaru, and S. O. Han. (2003) Cellulosomes from mesophilic bacteria. *J Bacteriol* **185:** 5907-14.

Edwards, U., T. Rogall, H. Blocker, M. Emde, and E. C. Bottger. (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* **17:** 7843-53.

Egan, S., S. James, and S. Kjelleberg. (2002) Identification and characterization of a putative transcriptional regulator controlling the expression of fouling inhibitors in Pseudoalteromonas tunicata. *Applied and Environmental Microbiology* **68:** 372-378.

Ehrlich, H., M. Krautter, T. Hanke, P. Simon, C. Knieb, S. Heinemann, and H. Worch. (2007) First evidence of the presence of chitin in skeletons of marine sponges. *J Exp Zoolog B Mol Dev Evol* **308:** 473-83.

Eilers, H., J. Pernthaler, F. O. Glockner, and R. Amann. (2000) Culturability and In situ abundance of pelagic bacteria from the North Sea. *Appl Environ Microbiol* **66:** 3044-51.

Ekborg, N. A., J. M. Gonzalez, M. B. Howard, L. E. Taylor, S. W. Hutcheson, and R. M. Weiner. (2005) *Saccharophagus degradans* gen. nov., sp. nov., a versatile marine degrader of complex polysaccharides. *Int J Syst Evol Microbiol* **55:** 1545-9.

Ekborg, N. A., W. Morrill, A. M. Burgoyne, L. Li, and D. L. Distel. (2007) CelAB, a multifunctional cellulase encoded by *Teredinibacter turnerae* T7902T, a culturable symbiont isolated from the wood-boring marine bivalve *Lyrodus pedicellatus*. *Appl Environ Microbiol* **73:** 7785-8.

Elifantz, H., R. R. Malmstrom, M. T. Cottrell, and D. L. Kirchman. (2005) Assimilation of polysaccharides and glucose by major bacterial groups in the Delaware Estuary. *Appl Environ Microbiol* **71:** 7799-805.

Ellegren, H. (2008) Sequencing goes 454 and takes large-scale genomics into the wild. *Molecular Ecology* **17:** 1629-1631.

Eng, J. K., A. L. Mccormack, and J. R. Yates. (1994) An Approach to Correlate Tandem Mass-Spectral Data of Peptides with Amino-Acid-Sequences in a Protein Database. *Journal of the American Society for Mass Spectrometry* **5:** 976-989.

Faurobert, M., E. Pelpoir, and J. Chaib. (2007) Phenol extraction of proteins for proteomic studies of recalcitrant plant tissues. *Methods Mol Biol* **355:** 9-14.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong, and C. M. Whitehouse. (1989) Electrospray ionization for mass spectrometry of large biomolecules. *Science* **246:** 64-71.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al. (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* **269:** 496-512.

Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W. Chisholm, and E. F. Delong. (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci* **105:** 3805-10.

Fuhrman, J. H., A. (2008) Bacterial and Archaeal Community Structure and it's Patterns, *In* Microbial Ecology in the Oceans, Second ed. D. L. Kirchman (ed.), Wiley-Blackwell. p 45-80.

Gabor, E. M., E. J. de Vries, and D. B. Janssen. (2003) Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. *Fems Microbiology Ecology* **44:** 153-163.

Gade, D., J. Gobom, and R. Rabus. (2005) Proteomic analysis of carbohydrate catabolism and regulation in the marine bacterium Rhodopirellula baltica. *Proteomics* **5:** 3672-83.

Gal, L., S. Pages, C. Gaudin, A. Belaich, C. Reverbel-Leroy, C. Tardif, and J. P. Belaich. (1997) Characterization of the cellulolytic complex (cellulosome) produced by Clostridium cellulolyticum. *Appl Environ Microbiol* **63:** 903-9.

Garsoux, G., J. Lamotte, C. Gerday, and G. Feller. (2004) Kinetic and structural optimization to catalysis at low temperatures in a psychrophilic cellulase from the Antarctic bacterium Pseudoalteromonas haloplanktis. *Biochem J* **384:** 247-53.

Gavin, A. C., M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. M. Rick, A. M. Michon, C. M. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. A. Heurtier, R. R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga. (2000) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415:** 141-7.

Gilbert, J. A., M. Muhling, and I. Joint. (2008) A rare SAR11 fosmid clone confirming genetic variability in the 'Candidatus Pelagibacter ubique' genome. *ISME J* **2:** 790-3.

Gilkes, N. R., R. A. Warren, R. C. Miller, Jr., and D. G. Kilburn. (1988) Precise excision of the cellulose binding domains from two *Cellulomonas fimi* cellulases by a homologous protease and the effect on catalysis. *J Biol Chem* **263:** 10401-7.

Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345:** 60-3.

Giovannoni, S. J., and U. Stingl. (2005) Molecular diversity and ecology of microbial plankton. *Nature* **437:** 343-8.

Gonzalez, J. M., B. Fernandez-Gomez, A. Fernandez-Guerra, L. Gomez-Consarnau, O. Sanchez, M. Coll-Llado, J. Del Campo, L. Escudero, R. Rodriguez-Martinez, L. Alonso-Saez, M. Latasa, I. Paulsen, O. Nedashkovskaya, I. Lekunberri, J. Pinhassi, and C. Pedros-Alio. (2008) Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (Flavobacteria). *Proc Natl Acad Sci* **105:** 8724-9.

Gooday, G. W. (1990) The Ecology of Chitin Degradation. *Advances in Microbial Ecology* **11:** 387-430.

Gram, L., H. P. Grossart, A. Schlingloff, and T. Kiorboe. (2002) Possible quorum sensing in marine snow bacteria: production of acylated homoserine lactones by *Roseobacter* strains isolated from marine snow. *Appl Environ Microbiol* **68:** 4111-6.

Griffiths, R. I., A. S. Whiteley, A. G. O'Donnell, and M. J. Bailey. (2000) Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Appl Environ Microbiol* **66:** 5488-91.

Grisez, L., and F. Ollevier. (1995) Comparative Serology of the Marine Fish Pathogen *Vibrio anguillarum*. *Appl Environ Microbiol* **61:** 4367-4373.

Guillard, R. R., and J. H. Ryther. (1962) Studies of marine planktonic diatoms. *Can J Microbiol* **8:** 229-39.

Guo, B., X. L. Chen, C. Y. Sun, B. C. Zhou, and Y. Z. Zhang. (2009) Gene cloning, expression and characterization of a new cold-active and salt-tolerant endo-beta-1,4-xylanase from marine *Glaciecola mesophila* KMM 241. *Appl Microbiol Biotechnol.*

Hager, J. W. (2002) A new linear ion trap mass spectrometer. *Rapid Communications in Mass Spectrometry* **16:** 512-526.

Hall, N. (2007) Advanced sequencing technologies and their wider impact in microbiology. *J Exp Biol* **210:** 1518-25.

Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman. (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chem Biol* **5:** 245-9.

Hashimoto, H. (2006) Recent structural studies of carbohydrate-binding modules. *Cell Mol Life Sci* **63:** 2954-67.

Henrissat, B. (1991) A Classification of Glycosyl Hydrolases Based on Amino-Acid-Sequence Similarities. *Biochemical Journal* **280:** 309-316.

Heukshoven, J. D., R. (1985) Simplified method for silver staining of protein in polyacrylamide gels and the mechanism of silver staining
Electrophoresis **6:** 103-112.

Hilden, L., and G. Johansson. (2004) Recent developments on cellulases and carbohydrate-binding modules with cellulose affinity. *Biotechnol Lett* **26:** 1683-93.

Holt, J. G., and R. A. Lewin. (1968) *Herpetosiphon aurantiacus* gen. et sp. n., a new filamentous gliding organism. *J Bacteriol* **95:** 2407-8.

Honda, Y., H. Taniguchi, and M. Kitaoka. (2008) A reducing-end-acting chitinase from *Vibrio proteolyticus* belonging to glycoside hydrolase family 19. *Appl Microbiol Biotechnol* **78:** 627-34.

Howard, M. B., N. A. Ekborg, R. M. Weiner, and S. W. Hutcheson. (2003) Detection and characterization of chitinases and other chitin-modifying enzymes. *J Ind Microbiol Biotechnol* **30:** 627-35.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. (2007) MEGAN analysis of metagenomic data. *Genome Res* **17:** 377-86.

Huson, D. H., D. C. Richter, S. Mitra, A. F. Auch, and S. C. Schuster. (2009) Methods for comparative metagenomics. *BMC Bioinformatics* **10 Suppl 1:** 12.

Hutchison, C. A., and J. C. Venter. (2006) Single-cell genomics. *Nature Biotechnology* **24:** 657-658.

Iriberri, J., M. Unanue, I. Barcina, and L. Egea. (1987) Seasonal Variation in Population Density and Heterotrophic Activity of Attached and Free-Living Bacteria in Coastal Waters. *Appl Environ Microbiol* **53:** 2308-2314.

Irwin, D., D. H. Shin, S. Zhang, B. K. Barr, J. Sakon, P. A. Karplus, and D. B. Wilson. (1998) Roles of the catalytic domain and two cellulose binding domains of *Thermomonospora fusca E4* in cellulose hydrolysis. *J Bacteriol* **180:** 1709-14.

Itoh, Y., J. Watanabe, H. Fukada, R. Mizuno, Y. Kezuka, T. Nonaka, and T. Watanabe. (2006) Importance of Trp59 and Trp60 in chitin-binding, hydrolytic, and antifungal activities of *Streptomyces griseus* chitinase C. *Appl Microbiol Biotechnol* **72:** 1176-84.

Itoi, S., Y. Kanomata, Y. Koyama, K. Kadokura, S. Uchida, T. Nishio, T. Oku, and H. Sugita. (2007) Identification of a novel endochitinase from a marine bacterium *Vibrio proteolyticus* strain No. 442. *Biochimica et Biophysica Acta* **1774:** 1099-1107.

Jarvis, M. (2003) Chemistry: cellulose stacks up. *Nature* **426:** 611-2.

Jensen, P. R., C. A. Kauffman, and W. Fenical. (1996) High recovery of culturable bacteria from the surfaces of marine algae. *Marine Biology* **126:** 1-7.

Johansen, J. E., P. Nielsen, and C. Sjoholm. (1999) Description of *Cellulophaga baltica* gen. nov., sp. nov. and *Cellulophaga fucicola* gen. nov., sp. nov. and reclassification of *Cytophaga lytica* to *Cellulophaga lytica* gen. nov., comb. nov. Int *J Syst Bacteriol* **3:** 1231-40.

Junge, K., C. Krembs, J. Deming, A. Stierle, and H. Eicken. (2001) A microscopic approach to investigate bacteria under in situ conditions in sea-ice samples. *Annals of Glaciology*, Vol 33 **33:** 304-310.

Kan, J., T. E. Hanson, J. M. Ginter, K. Wang, and F. Chen. (2005) Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems* **1:** 7-20.

Kang, S. C., S. Park, and D. G. Lee. (1998) Isolation and characterization of a chitinase cDNA from the entomopathogenic fungus, *Metarhizium anisopliae*. *Fems Microbiology Letters* **165:** 267-271.

Karl, D. M. (2007) Microbial oceanography: paradigms, processes and promise. *Nat Rev Microbiol* **5:** 759-69.

Karner, M., and G. J. Herndl. (1992) Extracellular Enzymatic-Activity and Secondary Production in Free-Living and Marine-Snow-Associated Bacteria. *Marine Biology* **113:** 341-347.

Kataeva, I. A., R. D. Seidel, 3rd, A. Shah, L. T. West, X. L. Li, and L. G. Ljungdahl. (2002) The fibronectin type 3-like repeat from the *Clostridium thermocellum* cellobiohydrolase

CbhA promotes hydrolysis of cellulose by modifying its surface. *Appl Environ Microbiol* **68:** 4292-300.

Kato, N., T. Sato, C. Kato, M. Yajima, J. Sugiyama, T. Kanda, M. Mizuno, K. Nozaki, S. Yamanaka, and Y. Amano. (2007) Viability and cellulose synthesizing ability of Gluconacetobacter xylinus cells under high-hydrostatic pressure. *Extremophiles* **11:** 693-8.

Kavoosi, M., J. Meijer, E. Kwan, A. L. Creagh, D. G. Kilburn, and C. A. Haynes. (2004) Inexpensive one-step purification of polypeptides expressed in *Escherichia coli* as fusions with the family 9 carbohydrate-binding module of xylanase 10A from T-maritima. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **807:** 87-94.

Kelly, K. M., and A. Y. Chistoserdov. (2001) Phylogenetic analysis of the succession of bacterial communities in the Great South Bay (Long Island). *FEMS Microbiol Ecol* **35:** 85-95.

Keyhani, N. O., and S. Roseman. (1999) Physiological aspects of chitin catabolism in marine bacteria. *Biochim Biophys Acta* **1473:** 108-22.

Kim, S. J., C. M. Lee, B. R. Han, M. Y. Kim, Y. S. Yeo, S. H. Yoon, B. S. Koo, and H. K. Jun. (2008) Characterization of a gene encoding cellulase from uncultured soil bacteria. *FEMS Microbiol Lett* **282:** 44-51.

Kim, U. J., H. Shizuya, P. J. de Jong, B. Birren, and M. I. Simon. (1992) Stable propagation of cosmid sized human DNA inserts in an F factor based vector. *Nucleic Acids Res* **20:** 1083-5.

Kiorboe, T., K. Tang, H. P. Grossart, and H. Ploug. (2003) Dynamics of microbial communities on marine snow aggregates: Colonization, growth, detachment, and grazing mortality of attached bacteria. *Applied and Environmental Microbiology* **69:** 3036-3047.

Kirchman, D. L. (2008) Introduction and overview, *In* Microbial Ecology of the Oceans. D. L. Kirchman (ed.), Wiley-Blackwell. p. 1-26.

Kirchman, D. L., and J. White. (1999) Hydrolysis and mineralization of chitin in the Delaware Estuary. *Aquatic Microbial Ecology* **18:** 187-196.

Kirchner, M. (1995) Microbial Colonization of Copepod Body Surfaces and Chitin Degradation in the Sea. *Helgolander Meeresuntersuchungen* **49:** 201-212.

Klaassens, E. S., W. M. de Vos, and E. E. Vaughan. (2007) Metaproteomics approach to study the functionality of the microbiota in the human infant gastrointestinal tract. *Appl Environ Microbiol* **73:** 1388-92.

Kondo, R., K. J. Purdy, S. D. Q. Silva, and D. B. Nedwell. (2007) Spatial dynamics of sulphate-reducing bacterial compositions in sediment along a salinity gradient in a UK estuary. *Microbes and Environments* **22:** 11-19.

Krause, L., N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye. (2008) Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research* **36:** 2230-2239.

Kunin, V., A. Copeland, A. Lapidus, K. Mavromatis, and P. Hugenholtz. (2008) A bioinformatician's guide to metagenomics. *Microbiol Mol Biol Rev* **72:** 557-78.

Lacerda, C. M., L. H. Choe, and K. F. Reardon. (2007) Metaproteomic analysis of a bacterial community response to cadmium exposure. *J Proteome Res* **6:** 1145-52.

Laemmli, U. K. (1970) Cleavage of Structural Proteins during Assembly of Head of Bacteriophage-T4. *Nature* **227:** 680-685.

Lamed, R., E. Setter, and E. A. Bayer. (1983) Characterization of a cellulose-binding, cellulase-containing complex in *Clostridium thermocellum. J Bacteriol* **156:** 828-36.

Leveau, J. H., S. Gerards, W. de Boer, and J. A. van Veen. (2004) Phylogeny-function analysis of metagenomic libraries: screening for expression of ribosomal RNA genes by large-insert library fluorescent in situ hybridization (LIL-FISH). *Environ Microbiol* **6:** 990-8.

Li, M., I. Rosenshine, S. L. Tung, X. H. Wang, D. Friedberg, C. L. Hew, and K. Y. Leung. (2004) Comparative proteomic analysis of extracellular proteins of enterohemorrhagic and enteropathogenic Escherichia coli strains and their ihf and ler mutants. Applied and *Environmental Microbiology* **70:** 5274-5282.

Li, X., and S. Roseman. (2004) The chitinolytic cascade in Vibrios is regulated by chitin oligosaccharides and a two-component chitin catabolic sensor/kinase. *Proc Natl Acad Sci* **101:** 627-31.

Liu, J., M. J. McBride, and S. Subramaniam. (2007) Cell surface filaments of the gliding bacterium *Flavobacterium johnsoniae* revealed by cryo-electron tomography. *J Bacteriol* **189:** 7503-6.

Lo, I., V. J. Denef, N. C. Verberkmoes, M. B. Shah, D. Goltsman, G. DiBartolo, G. W. Tyson, E. E. Allen, R. J. Ram, J. C. Detter, P. Richardson, M. P. Thelen, R. L. Hettich, and J. F. Banfield. (2007) Strain-resolved community proteomics reveals recombining genomes of acidophilic bacteria. *Nature* **446:** 537-41.

Lorenz, P., and J. Eck. (2005) Metagenomics and industrial applications. *Nat Rev Microbiol* **3:** 510-6.

Margulies, M., M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y. J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. Alenquer, T. P. Jarvie, K. B. Jirage, J. B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, and J. M. Rothberg. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376-80.

Martinez, J.P & Gozalbo, D. (2001) Chitin. Encyclopedia of Life Sciences. John Wiley and Son's. P1-8.

Matsuyama, H., T. Hirabayashi, H. Kasahara, H. Minami, T. Hoshino, and I. Yumoto. (2006) *Glaciecola chathamensis* sp. nov., a novel marine polysaccharide-producing bacterium. *Int J Syst Evol Microbiol* **56:** 2883-6.

Mccarter, J. D., M. J. Adam, N. G. Hartman, and S. G. Withers. (1994) In-Vivo Inhibition of Beta-Glucosidase and Beta-Mannosidase Activity in Rats by 2-Deoxy-2-Fluoro-Beta-Glycosyl Fluorides and Recovery of Activity in-Vivo and in-Vitro. Biochemical Journal **301:** 343-348.

McCormack, A. L., D. M. Schieltz, B. Goode, S. Yang, G. Barnes, D. Drubin, and J. R. Yates. (1997) Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Analytical Chemistry* **69:** 767-776.

McDonald, J. E., A. B. de Menezes, H. E. Allison, and A. J. McCarthy. (2009) Molecular biological detection and quantification of novel *Fibrobacter* populations in freshwater lakes. *Appl Environ Microbiol* **75:** 5148-52.

Medini, D., D. Serruto, J. Parkhill, D. A. Relman, C. Donati, R. Moxon, S. Falkow, and R. Rappuoli. (2008) Microbiology in the post-genomic era. *Nat Rev Microbiol* **6:** 419-30.

Meng, J., F. Wang, Y. Zheng, X. Peng, H. Zhou, and X. Xiao. (2009) An uncultivated *Crenarchaeota* contains functional bacteriochlorophyll a synthase. *ISME J* **3:** 106-16.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386.

Miyamoto, K., E. Nukui, H. Itoh, T. Sato, T. Kobayashi, C. Imada, E. Watanabe, Y. Inamori, and H. Tsujibo. (2002) Molecular analysis of the gene encoding a novel chitin-binding protease from *Alteromonas sp.* strain O-7 and its role in the chitinolytic system. *J Bacteriol* **184:** 1865-72.

Moran, M. A. (2008) Genomics and Metagenomics of marine prokaryotes, *In* Microbial Ecology in the Oceans, Second ed. D. L. Kirchman (ed.), Wiley-Blackwell. p. 91-130,

Mou, X. Z., S. L. Sun, R. A. Edwards, R. E. Hodson, and M. A. Moran. (2008) Bacterial carbon processing by generalist species in the coastal ocean. *Nature* **451:** 708-U4.

Nagata. (2008) Organic Matter-Bacteria Interactions in Seawater. *In* Microbial Ecology in the Oceans, Second ed. D. L. Kirchman (ed.), Wiley-Blackwell. p. 207-242

Neena Din, N. R. G., Bahar Tekant, Robert C. Miller Jr., R. Anthony J. Warren & Douglas G. Kilburn. (1991) Non–Hydrolytic Disruption of Cellulose Fibres by the Binding Domain of a Bacterial Cellulase. *Nature Biotechnology* **9:** 1096-1099.

Neufeld, J. D., Y. Chen, M. G. Dumont, and J. C. Murrell. (2008) Marine methylotrophs revealed by stable-isotope probing, multiple displacement amplification and metagenomics. *Environ Microbiol* **10:** 1526-35.

Nogales, B., K. N. Timmis, D. B. Nedwell, and A. M. Osborn. (2002) Detection and diversity of expressed denitrification genes in estuarine sediments after reverse transcription-PCR amplification from mRNA. *Appl Environ Microbiol* **68:** 5017-25.

O'Farrell, P. H. (1975) High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250:** 4007-21.

Oda, K., D. Kakizono, O. Yamada, H. Iefuji, O. Akita, and K. Iwashita. (2006) Proteomic analysis of extracellular proteins from *Aspergillus oryzae* grown under submerged and solid-state culture conditions. *Appl Environ Microbiol* **72:** 3448-57.

Ohara, H., S. Karita, T. Kimura, K. Sakka, and K. Ohmiya. (2000) Characterization of the cellulolytic complex (cellulosome) from *Ruminococcus albus*. *Biosci Biotechnol Biochem* **64:** 254-60.

Ohishi, K., K. Murase, T. Ohta, and H. Etoh. (2000) Cloning and sequencing of a chitinase gene from *Vibrio alginolyticus* H-8. *J Biosci Bioeng* **89:** 501-5.

Olsen, J. V., and M. Mann. (2004) Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation. *Proc Natl Acad Sci* **101:** 13417-22.

Overbeek, R., T. Begley, R. M. Butler, J. V. Choudhuri, H. Y. Chuang, M. Cohoon, V. de Crecy-Lagard, N. Diaz, T. Disz, R. Edwards, M. Fonstein, E. D. Frank, S. Gerdes, E. M. Glass, A. Goesmann, A. Hanson, D. Iwata-Reuyl, R. Jensen, N. Jamshidi, L. Krause, M. Kubal, N. Larsen, B. Linke, A. C. McHardy, F. Meyer, H. Neuweger, G. Olsen, R. Olson, A. Osterman, V. Portnoy, G. D. Pusch, D. A. Rodionov, C. Ruckert, J. Steiner, R. Stevens, I. Thiele, O. Vassieva, Y. Ye, O. Zagnitko, and V. Vonstein. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Research* **33:** 5691-5702.

Pang, H., P. Zhang, C. J. Duan, X. C. Mo, J. L. Tang, and J. X. Feng. (2009) Identification of cellulase genes from the metagenomes of compost soils and functional characterization of one novel endoglucanase. *Curr Microbiol* **58:** 404-8.

Pantoom, S., C. Songsiriritthigul, and W. Suginta. (2008) The effects of the surface-exposed residues on the binding and hydrolytic activities of *Vibrio carchariae* chitinase A. BMC Biochem **9:** 2-13.

Park, C., J. T. Novak, R. F. Helm, Y. O. Ahn, and A. Esen. (2008) Evaluation of the extracellular proteins in full-scale activated sludges. *Water Res* **42:** 3879-89.

Park, S. J., C. H. Kang, J. C. Chae, and S. K. Rhee. (2008) Metagenome microarray for screening of fosmid clones containing specific genes. *Fems Microbiology Letters* **284:** 28-34.

Pedraza-Reyes, M., and F. Gutierrez-Corona. (1997) The bifunctional enzyme chitosanase-cellulase produced by the gram-negative microorganism *Myxobacter sp.* AL-1 is highly similar to *Bacillus subtilis* endoglucanases. *Arch Microbiol* **168:** 321-7.

Penesyan, A., Z. Marshall-Jones, C. Holmstrom, S. Kjelleberg, and S. Egan. (2009) Antimicrobial activity observed among cultured marine epiphytic bacteria reflects their potential as a source of new drugs. *Fems Microbiology Ecology* **69:** 113-124.

Phillips, C. I., and M. Bogyo. (2005) Proteomics meets microbiology: technical advances in the global mapping of protein expression and function. *Cell Microbiol* **7:** 1061-76.

Pierre-Alain, M., M. Christophe, S. Severine, A. Houria, L. Philippe, and R. Lionel. (2007) Protein extraction and fingerprinting optimization of bacterial communities in natural environment. *Microb Ecol* **53:** 426-34.

Ponpium, P., K. Ratanakhanokchai, and K. L. Kyu. (2000) Isolation and properties of a cellulosome-type multienzyme complex of the thermophilic *Bacteroides* sp. strain P-1. *Enzyme Microb Technol* **26:** 459-465.

Powell, M. J., J. N. Sutton, C. E. Del Castillo, and A. I. Timperman. (2005) Marine proteomics: generation of sequence tags for dissolved proteins in seawater using tandem mass spectrometry. *Marine Chemistry* **95:** 183-198.

Prabagaran, S. R., R. Manorama, D. Delille, and S. Shivaji. (2007) Predominance of *Roseobacter, Sulfitobacter, Glaciecola* and *Psychrobacter* in seawater collected off Ushuaia, Argentina, Sub-Antarctica. *FEMS Microbiol Ecol* **59:** 342-55.

Pugsley, A. P., C. Chapon, and M. Schwartz. (1986) Extracellular pullulanase of *Klebsiella pneumoniae* is a lipoprotein. *J Bacteriol* **166:** 1083-8.

Purdy, K. J., M. A. Munson, D. B. Nedwell, and T. Martin Embley. (2002) Comparison of the molecular diversity of the methanogenic community at the brackish and marine ends of a UK estuary. *FEMS Microbiol Ecol* **39:** 17-21.

Qi, W., G. Nong, J. F. Preston, F. Ben-Ami, and D. Ebert. (2009) Comparative metagenomics of Daphnia symbionts. *BMC Genomics* **10:** 172.

Raes, J., K. U. Foerstner, and P. Bork. (2007) Get the most out of your metagenome: computational analysis of environmental sequence data. *Curr Opin Microbiol* **10:** 490-8.

Ram, R. J., N. C. Verberkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, 2nd, M. Shah, R. L. Hettich, and J. F. Banfield. (2005) Community proteomics of a natural microbial biofilm. *Science* **308:** 1915-20.

Rappe, M. S., S. A. Connon, K. L. Vergin, and S. J. Giovannoni. (2002) Cultivation of the ubiquitous SAR11 marine bacterioplankton clade. *Nature* **418:** 630-3.

Rath, J., K. Y. Wu, G. J. Herndl, and E. F. DeLong. (1998) High phylogenetic diversity in a marine-snow-associated bacterial assemblage. *Aquatic Microbial Ecology* **14:** 261-269.

Ravikumar, M. N. V. (1999) Chitin and chitosan fibres: A review. Bulletin of Materials *Science* **22:** 905-915.

Reynolds, J. A., and C. Tanford. (1970) The gross conformation of protein-sodium dodecyl sulfate complexes. *J Biol Chem* **245:** 5161-5.

Riaz, K., C. Elmerich, D. Moreira, A. Raffoux, Y. Dessaux, and D. Faure. (2008) A metagenomic analysis of soil bacteria extends the diversity of quorum-quenching lactonases. *Environ Microbiol* **10:** 560-70.

Riemann, L., and F. Azam. (2002) Widespread N-acetyl-D-glucosamine uptake among pelagic marine bacteria and its ecological implications. *Appl Environ Microbiol* **68:** 5554-62.

Riesenfeld, C. S., P. D. Schloss, and J. Handelsman. (2004) Metagenomics: genomic analysis of microbial communities. *Annu Rev Genet* **38:** 525-52.

Rodriguez, B., M. Kavoosi, J. Koska, A. L. Creagh, D. G. Kilburn, and C. A. Haynes. (2004) Inexpensive and generic affinity purification of recombinant proteins using a family 2a CBM fusion tag. *Biotechnology Progress* **20:** 1479-1489.

Roling, F. M. H., I. M. (2005) Prokaryotic systematics: PCR and sequence analysis of amplified 16S rRNA genes. *In* Molecular Microbial Ecology. Osborn, M. A. & Smith, C.J. (ed.), Taylor & Francis Group, p. 25-57.

Romanenko, L. A., N. V. Zhukova, M. Rohde, A. M. Lysenko, V. V. Mikhailov, and E. Stackebrandt. (2003) *Glaciecola mesophila* sp. nov., a novel marine agar-digesting bacterium. *Int J Syst Evol Microbiol* **53:** 647-51.

Ronaghi, M., S. Karamohamed, B. Pettersson, M. Uhlen, and P. Nyren. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem* **242:** 84-9.

Rothberg, J. M., and J. H. Leamon. (2008) The development and impact of 454 sequencing. *Nat Biotechnol* **26:** 1117-24.

Rusch, D. B., A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcon, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C.

Venter. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. PLoS Biol **5:** e77.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream, and B. Barrell. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16:** 944-5.

Sabathe, F., A. Belaich, and P. Soucaille. (2002) Characterization of the cellulolytic complex (cellulosome) of *Clostridium acetobutylicum*. *FEMS Microbiol Lett* **217:** 15-22.

Sambrook & Russell (2001) Molecular Cloning: A Laboratory Manual. Third Edition. Cold Spring Harbour Laboratory Press.

Sanger, F., S. Nicklen, and A. R. Coulson. (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci* **74:** 5463-7.

Schwarz, W. H. (2001) The cellulosome and cellulose degradation by anaerobic bacteria. Appl Microbiol Biotechnol **56:** 634-49.

Schweder, T., S. Markert, and M. Hecker. (2008) Proteomics of marine bacteria. *Electrophoresis* **29:** 2603-16.

Seydel, A., P. Gounon, and A. P. Pugsley. (1999) Testing the '+2 rule' for lipoprotein sorting in the Escherichia coli cell envelope with a new genetic selection. *Mol Microbiol* **34:** 810-21.

Shapiro, A. L., E. Vinuela, and J. V. Maizel, Jr. (1967) Molecular weight estimation of polypeptide chains by electrophoresis in SDS-polyacrylamide gels. *Biochem Biophys Res Commun* **28:** 815-20.

Shizuya, H., B. Birren, U. J. Kim, V. Mancino, T. Slepak, Y. Tachiiri, and M. Simon. (1992) Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci* **89:** 8794-7.

Shoham, Y., R. Lamed, and E. A. Bayer. (1999) The cellulosome concept as an efficient microbial strategy for the degradation of insoluble polysaccharides. *Trends Microbiol* **7:** 275-81.

Shoseyov, O., Z. Shani, and I. Levy. (2006) Carbohydrate binding modules: Biochemical properties and novel applications. Microbiology and Molecular Biology Reviews **70:** 283-295

Simon, M., H. P. Grossart, B. Schweitzer, and H. Ploug. (2002) Microbial ecology of organic aggregates in aquatic ecosystems. *Aquatic Microbial Ecology* **28:** 175-211.

Simpson, P. J., H. Xie, D. N. Bolam, H. J. Gilbert, and M. P. Williamson. (2000) The structural basis for the ligand specificity of family 2 carbohydrate-binding modules. *J Biol Chem* **275:** 41137-42.

Simpson, R. J. (2003) Proteins and Proteomics: A Laboratory manual Cold Springs Harbour Laboratory Press.

Singla, A. K., and M. Chawla. (2001) Chitosan: some pharmaceutical and biological aspects-an update. *J Pharm Pharmacol* **53:** 1047-67.

Siuti, N., and N. L. Kelleher. (2007) Decoding protein modifications using top-down mass spectrometry. *Nat Methods* **4:** 817-21.

Skovhus, T. L., C. Holmstrom, S. Kjelleberg, and I. Dahllof. (2007) Molecular investigation of the distribution, abundance and diversity of the genus *Pseudoalteromonas* in marine samples. *FEMS Microbiol Ecol* **61:** 348-61.

Smelcerovic, A., Z. Knezevic-Jugovic, and Z. Petronijevic. (2008) Microbial polysaccharides and their derivatives as current and prospective pharmaceuticals. *Curr Pharm Des* **14:** 3168-95.

Smith, C. J., D. B. Nedwell, L. F. Dong, and A. M. Osborn. (2007) Diversity and abundance of nitrate reductase genes (narG and napA), nitrite reductase genes (nirS and nrfA), and their transcripts in estuarine sediments. *Appl Environ Microbiol* **73:** 3612-22.

Smith, D. C., M. Simon, A. L. Alldredge, and F. Azam. (1992) Intense Hydrolytic Enzyme-Activity on Marine Aggregates and Implications for Rapid Particle Dissolution. *Nature* **359:** 139-142.

Staden, R. (1996) The Staden sequence analysis package. *Molecular Biotechnology* **5:** 233-241.

Standing, K. G. (2003) Peptide and protein *de novo* sequencing by mass spectrometry. *Curr Opin Struct Biol* **13:** 595-601.

Steele, H. L., and W. R. Streit. (2005) Metagenomics. *advances in ecology and biotechnology. FEMS Microbiol Lett* **247:** 105-11.

Steenbakkers, P. J., A. Freelove, B. Van Cranenbroek, B. M. Sweegers, H. R. Harhangi, G. D. Vogels, G. P. Hazlewood, H. J. Gilbert, and H. J. Op den Camp. (2002) The major component of the cellulosomes of anaerobic fungi from the genus *Piromyces* is a family 48 glycoside hydrolase. *DNA Seq* **13:** 313-20.

Stingl, U., H. J. Tripp, and S. J. Giovannoni. (2007) Improvements of high-throughput culturing yielded novel SAR11 strains and other abundant marine bacteria from the Oregon coast and the Bermuda Atlantic Time Series study site. *ISME J* **1:** 361-71.

Suenaga, H., T. Ohnuki, and K. Miyazaki. (2007) Functional screening of a metagenomic library for genes involved in microbial degradation of aromatic compounds. *Environmental Microbiology* **9:** 2289-2297.

Sunna, A., M. Moracci, M. Rossi, and G. Antranikian. (1997) Glycosyl hydrolases from hyperthermophiles. *Extremophiles* **1:** 2-13.

Suttle, C. A. 2007. Marine viruses--major players in the global ecosystem. *Nat Rev Microbiol* **5:** 801-12.

Svitil, A. L., S. Chadhain, J. A. Moore, and D. L. Kirchman. (1997) Chitin Degradation Proteins Produced by the Marine Bacterium *Vibrio harveyi* Growing on Different Forms of Chitin. *Appl Environ Microbiol* **63:** 408-413.

Tabb, D. L., W. H. McDonald, and J. R. Yates, 3rd. (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **1:** 21-6.

Tanaka, T., S. Fujiwara, S. Nishikori, T. Fukui, M. Takagi, and T. Imanaka. (1999) A unique chitinase with dual active sites and triple substrate binding sites from the hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. *Appl Environ Microbiol* **65:** 5338-44.

Taylor, L. E., 2nd, B. Henrissat, P. M. Coutinho, N. A. Ekborg, S. W. Hutcheson, and R. M. Weiner. (2006) Complete cellulase system in the marine bacterium *Saccharophagus degradans* strain 2-40T. *J Bacteriol* **188:** 3849-61.

Thomas, T., F. F. Evans, D. Schleheck, A. Mai-Prochnow, C. Burke, A. Penesyan, D. S. Dalisay, S. Stelzer-Braid, N. Saunders, J. Johnson, S. Ferriera, S. Kjelleberg, and S. Egan. (2008) Analysis of the *Pseudoalteromonas tunicata* genome reveals properties of a surface-associated life style in the marine environment. *PLoS ONE* **3:** e3252.

Tomme, P., E. Kwan, N. R. Gilkes, D. G. Kilburn, and R. A. Warren. (1996) Characterization of CenC, an enzyme from *Cellulomonas fimi* with both endo- and exoglucanase activities. *J Bacteriol* **178:** 4216-23.

Tomme, P., H. Van Tilbeurgh, G. Pettersson, J. Van Damme, J. Vandekerckhove, J. Knowles, T. Teeri, and M. Claeyssens. (1988) Studies of the cellulolytic system of *Trichoderma reesei* QM 9414. Analysis of domain function in two cellobiohydrolases by limited proteolysis. *Eur J Biochem* **170:** 575-81.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554-7.

Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428:** 37-43.

Urich, T., A. Lanzen, J. Qi, D. H. Huson, C. Schleper, and S. C. Schuster. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* **3:** e2527.

Van Trappen, S., T. L. Tan, J. Yang, J. Mergaert, and J. Swings. (2004) *Glaciecola polaris* sp. nov., a novel budding and prosthecate bacterium from the Arctic Ocean, and emended description of the genus *Glaciecola*. *Int J Syst Evol Microbiol* **54:** 1765-71.

Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66-74.

VerBerkmoes, N. C., H. M. Connelly, C. Pan, and R. L. Hettich. (2004) Mass spectrometric approaches for characterizing bacterial proteomes. *Expert Rev Proteomics* **1:** 433-47.

VerBerkmoes, N. C., V. J. Denef, R. L. Hettich, and J. F. Banfield. (2009) Systems biology: Functional analysis of natural microbial consortia using community proteomics. *Nat Rev Microbiol* **7:** 196-205.

Verberkmoes, N. C., A. L. Russell, M. Shah, A. Godzik, M. Rosenquist, J. Halfvarson, M. G. Lefsrud, J. Apajalahti, C. Tysk, R. L. Hettich, and J. K. Jansson. (2009) Shotgun metaproteomics of the human distal gut microbiota. *ISME J* **3:** 179-89.

Violot, S., N. Aghajari, M. Czjzek, G. Feller, G. K. Sonan, P. Gouet, C. Gerday, R. Haser, and V. Receveur-Brechot. (2005) Structure of a full length psychrophilic cellulase from *Pseudoalteromonas haloplanktis* revealed by X-ray diffraction and small angle X-ray scattering. *Journal of Molecular Biology* **348:** 1211-1224.

Voget, S., H. L. Steele, and W. R. Streit. (2006) Characterization of a metagenome-derived halotolerant cellulase. *J Biotechnol* **126:** 26-36.

Warnecke, F., P. Luginbuhl, N. Ivanova, M. Ghassemian, T. H. Richardson, J. T. Stege, M. Cayouette, A. C. McHardy, G. Djordjevic, N. Aboushadi, R. Sorek, S. G. Tringe, M. Podar, H. G. Martin, V. Kunin, D. Dalevi, J. Madejska, E. Kirton, D. Platt, E. Szeto, A. Salamov, K. Barry, N. Mikhailova, N. C. Kyrpides, E. G. Matson, E. A. Ottesen, X. Zhang, M. Hernandez, C. Murillo, L. G. Acosta, I. Rigoutsos, G. Tamayo, B. D. Green, C. Chang, E. M. Rubin, E. J. Mathur, D. E. Robertson, P. Hugenholtz, and J. R. Leadbetter. (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450:** 560-5.

Watanabe, T., M. Kono, K. Aida, and H. Nagasawa. (1998) Purification and molecular cloning of a chitinase expressed in the hepatopancreas of the penaeid prawn *Penaeus japonicus*. *Biochim Biophys Acta* **1382:** 181-5.

Weiner, R. M., L. E. Taylor, 2nd, B. Henrissat, L. Hauser, M. Land, P. M. Coutinho, C. Rancurel, E. H. Saunders, A. G. Longmire, H. Zhang, E. A. Bayer, H. J. Gilbert, F. Larimer, I. B. Zhulin, N. A. Ekborg, R. Lamed, P. M. Richardson, I. Borovok, and S. Hutcheson. (2008) Complete genome sequence of the complex carbohydrate-degrading marine bacterium, *Saccharophagus degradans* strain 2-40 T. *PLoS Genet* **4:** 1000087.

Wery, N., U. Gerike, A. Sharman, J. B. Chaudhuri, D. W. Hough, and M. J. Danson. (2003) Use of a packed-column bioreactor for isolation of diverse protease-producing bacteria from antarctic soil. *Appl Environ Microbiol* **69:** 1457-64.

Wilder, M. N., N. Fusetani, and K. Aida. (1995) The Presence of 20-Hydroxyecdysonoic Acid and Ecdysonoic Acid in Eggs of the Giant Fresh-Water Prawn *Macrobrachium-Rosenbergii*. *Fisheries Science* **61:** 101-106.

Wilmes, P., A. F. Andersson, M. G. Lefsrud, M. Wexler, M. Shah, B. Zhang, R. L. Hettich, P. L. Bond, N. C. VerBerkmoes, and J. F. Banfield. (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. *ISME J* **2:** 853-64.

Wilmes, P., and P. L. Bond. (2004) The application of two-dimensional polyacrylamide gel electrophoresis and downstream analyses to a mixed community of prokaryotic microorganisms. *Environ Microbiol* **6:** 911-20.

Wilmes, P., M. Wexler, and P. L. Bond. (2008) Metaproteomics provides functional insight into activated sludge wastewater treatment. *PLoS ONE* **3:** 1778.

Wilson, D. B. (2008) Three microbial strategies for plant cell wall degradation. *Ann N Y Acad Sci* **1125:** 289-97.

Woebken, D., H. Teeling, P. Wecker, A. Dumitriu, I. Kostadinov, E. F. Delong, R. Amann, and F. O. Glockner. (2007) Fosmids of novel marine *Planctomycetes* from the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. *ISME J* **1:** 419-35.

Wommack, K. E., J. Bhavsar, and J. Ravel. (2008) Metagenomics: read length matters. *Appl Environ Microbiol* **74:** 1453-63.

Woyke, T., G. Xie, A. Copeland, J. M. Gonzalez, C. Han, H. Kiss, J. H. Saw, P. Senin, C. Yang, S. Chatterji, J. F. Cheng, J. A. Eisen, M. E. Sieracki, and R. Stepanauskas. (2009) Assembling the marine metagenome, one cell at a time. *PLoS ONE* **4:** e5299.

Xie, G., D. C. Bruce, J. F. Challacombe, O. Chertkov, J. C. Detter, P. Gilna, C. S. Han, S. Lucas, M. Misra, G. L. Myers, P. Richardson, R. Tapia, N. Thayer, L. S. Thompson, T. S. Brettin, B. Henrissat, D. B. Wilson, and M. J. McBride. (2007) Genome sequence of the cellulolytic gliding bacterium *Cytophaga hutchinsonii*. *Appl Environ Microbiol* **73:** 3536-46.

Xiong, P. J., and J. J. Wen. (2004) [Characterization and gene cloning of the endoglucanase from *Pseudoalteromonas* sp. DY3 strain]. *Chinese Journal of Biotechnology* **20:** 233-7.

Xu, C. G., X. J. Fan, Y. J. Fu, and A. H. Liang. (2008) Effect of location of the His-tag on the production of soluble and functional Buthus martensii Karsch insect toxin. *Protein Expr Purif* **59:** 103-9.

Yang, J. C., R. Madupu, A. S. Durkin, N. A. Ekborg, C. S. Pedamallu, J. B. Hostetler, D. Radune, B. S. Toms, B. Henrissat, P. M. Coutinho, S. Schwarz, L. Field, A. E. Trindade-Silva, C. A. Soares, S. Elshahawi, A. Hanora, E. W. Schmidt, M. G. Haygood, J. Posfai, J. Benner, C. Madinger, J. Nove, B. Anton, K. Chaudhary, J. Foster, A. Holman, S. Kumar, P. A. Lessard, Y. A. Luyten, B. Slatko, N. Wood, B. Wu, M. Teplitski, J. D. Mougous, N. Ward, J. A. Eisen, J. H. Badger, and D. L. Distel. (2009) The complete genome of *Teredinibacter turnerae* T7901: an intracellular endosymbiont of marine wood-boring bivalves (shipworms). PLoS ONE **4:** e6085.

Yates, J. R., 3rd. (2004) Mass spectral analysis in proteomics. *Annu Rev Biophys Biomol Struct* **33:** 297-316.

Yokobata, K., B. Trenchak, and P. J. de Jong. (1991) Rescue of unstable cosmids by in vitro packaging. *Nucleic Acids Res* **19:** 403-4.

Yong, J. J., S. J. Park, H. J. Kim, and S. K. Rhee. (2007) *Glaciecola agarilytica* sp. nov., an agar-digesting marine bacterium from the East Sea, Korea. *Int J Syst Evol Microbiol* **57:** 951-3.

Yoshikoshi, K., and Y. Ko. 1988. Structure and Function of the Peritrophic Membranes of Copepods. *Nippon Suisan Gakkaishi* **54:** 1077-1082.

Yu, C., A. M. Lee, B. L. Bassler, and S. Roseman. (1991) Chitin utilization by marine bacteria. A physiological function for bacterial adhesion to immobilized carbohydrates. *J Biol Chem* **266:** 24260-7.

Zeng, R., P. Xiong, and J. Wen. (2006) Characterization and gene cloning of a cold-active cellulase from a deep-sea psychrotrophic bacterium *Pseudoalteromonas sp.* DY3. Extremophiles **10:** 79-82.

Zhang, D. C., Y. Yu, B. Chen, H. X. Wang, H. C. Liu, X. Z. Dong, and P. J. Zhou. (2006) *Glaciecola psychrophila* sp. nov., a novel psychrophilic bacterium isolated from the Arctic. *Int J Syst Evol Microbiol* **56:** 2867-9.

Zhang, K., A. C. Martiny, N. B. Reppas, K. W. Barry, J. Malek, S. W. Chisholm, and G. M. Church. (2006) Sequencing genomes from single cells by polymerase cloning. *Nat Biotechnol* **24:** 680-6.

Zhang, Y. H., and L. R. Lynd. (2004) Toward an aggregated understanding of enzymatic hydrolysis of cellulose: noncomplexed cellulase systems. *Biotechnol Bioeng* **88:** 797-824.

Zhou, J., and D. K. Thompson. (2002) Challenges in applying microarrays to environmental studies. *Curr Opin Biotechnol* **13:** 204-7.

Zobell, C. E. (1941) Studies on marine bacteria. The cultural requirements of heterotrophic aerobes. *Journal of Marine Research* **4:** 42-75.

Appendix A: Amino acid sequences of ORF 9, 10, 11 and 13 as predicted by Artemis (Rutherford *et al.*, 2000)

ORF 9 predicted amino acid sequence

MSNRNLRSFIGPISRIMLLGLVFLLPIFMFIAAPQHAQAVAPRATTVACSVNYDVVNQWG
SGFQVNVTVTNNTTTAVDGWNLTWTFDQGQAFGSGWNASFAPSGSNMSASNTAGHWNGTI
GANGGTVAFGFQGTTTGGVTVPTNFAVNGVSCLDDVVPTVVPTAEPTDEPTVVPTDEPTV
VPTDEPTVVPTDEPTVVPTTNPGTDSCSVDYVLGNQWNTGFQADVTITNNSNSAIQGWNL
VWTFGGNEQFDSGWNATFSKAGQTVTVSNPASAWNGTIAANGGTASFGFQGTHSGNVVVP
ATFTLNGDVCGDGITPTVEPTVDPTVEPTVDPTIEPTVDPTVEPTVPPVTPVPGEHVQNP
FVGADAYLNPDYTEKVETQAATVGGTLGAQMELVAQYPSAVWMDRIGAITGSDSVMSLEA
HLDAALAQQNGDTPMTIMVVVYDLPNRDCSSEASNGELRVEEDGLNRYKTEYIDAIYAIF
AQPQYSDLRIVVILEPDSLPNLVTNLSIPDCQQAQTAYVEGVQYAISTLHPIDNVYIYLD
IAHSGWLGWPNNFDAAITLYTSTIAATVDGLNSIDGFVSNTSNTTPVIEPFLPDDTLVIP
GENLPIRSADIYEWNPYFSELPFVTAFRDSLIANGFPSDIGMLIDTSRNGWGGPDRPTAV
STATNLNDYVDESRIDTRPHRGGWCNQVGAGIGERPIAAPAPGVDAYVWVKPPGESDGIS
DPNFEVDPDDPAKQHDPMCDPNAQSVYNSAVPTNALDGAPHAGRWFAAQFQALVENAYPP
LE

ORF 10 predicted amino acid sequence

MKTEQEFIKKGLVLGVSLLLMAIIIMSSTPKTSANSANPYLWPYNQSTNISFNESDVYDA
WTAWRDAQITSNNAGGNGRYRVMGGVDGGSTVSEGQAYGILYTSIFDEQTLFDGLFLFAK
DHYNTQGVMDWHIGSPGVRIGSGGATDAEVDMAMGLVNACVKVQQNAWSASSAGIDYCVE
ATNLINAIYTYEVDHAGSSPPGGLPNNQGNELLPGDTWDVSGTYPDGIINLSYFPPGYFT
VFGKFTQNEAAWNAVIDRNYEVTDLVQAKSDNCSGLVPNWNKYNGDAQLVSWQTNNYSWW
SYDAARFAWRIAVDQAWYGRPEATETMNEIGGFFSSTGFNNIGEHSMNGIKTGSGPWPFF
VANAASAVWAAPNPVATNCGTGTGSLQESPQSAYNRVLSTKDNPNSYYGNAWRLFSMLLM
TGNFPNFYEMADGNVTPVPTSTPGSATNTPVPPTATIPAGTGACHVDYVVANEWGSGFQA
NVTITNNMSSAIDGYTLTWTHAPGQVVSSGWNVTVSQTGNQVTATNPAGSWMGKINANGG
TSSFGFQGSLTSKAVVPTDFVLNGTACNGDTPPTETPIPTPETPTVTPTVWCPQATSVPL
VVEPVTSPTNELSQTLVVKVYADWVSATGPAGSVTVDTPEADGFHVTVPLAANSINNISV
KSQIPVVTNPNGCTYGGYTLSKTVTIVQESDAVTPTLTPTATATATATPTATATTPSGTA
TCSVAYTVGNDWGSGFTTDVKITNKGASTINGWTLTYTYAGNQTITNAWNATVTQSGKTI
TATDAGWNGTLPPNGSASFGFQGSYSGSNIAPTTFKVNGSVCQ

ORF 11 predicted amino acid sequence

MSVQDEFLSAFFSKSKHFWRNILKQHNKRSTSFTFAALSIAFVLMVALFSFGSMATLPVA
AQESCSVDYVIVNQWNSGFQANVVITNNGSTPVNGWDLTWTLGSGQQFGSGWNATFASTG
SSVSASNVASHWNGTIGANGGTAAFGFQGTKGSGSATVPTDFAVNGVVCSGDIPPTATNV
PPTATDVGPTATDVPPTATDIPPTATDIPPTATDVPPTVTTIPPTATDVPPAGSCSVDYT
IANQWGSGFQGNVTITNHNTTAVSGYTLTWDFTNGQQLDSGWNATFSQTGTAVSVSNPAS
NWNGTINPNGTSSFGFQASLNGSNPIPTNFALNGEACGGGPEPTPEPTVGPTSTPGPSPT
PAPAALFRVNTEGRITKDGEVLPVQCGSWFGLEGRHEPSNDPINPSGAAMELYVGNTSWG
NGGGRTIQQTMDEITAMGINVVRMPVSPQTLDPTDPQGMAPNLKNHESVRVPNARQALEE
FIVLADQNNIEVMLDMHSCSNYLGWRAGRLDARPPYADWDRDLYDYYREDSSCAETLNPD
GVTRIQAYDETKWLNDLRTLAGMGQDLGVDNIIGIDIFNEPWDYTWEEWKTLSEHAYEAI
NEVNPNTLVFVQGISATADNQDGSPETITEVPHGDPATNPNWGENLFEAGTNIPNIPKDQ
LVYSPHTYGPSVFVQKGFMDPAQPECAGLEGDAAGDADCNIVINVPQLRSGWEEHFGYLK
DQGYAIVVGEFGGNLDWPLGGASLRDQGRWSHITPGVDQVWQDAFVDYMVEKGIEGCYWS
INPESGDTGGWYGHAYDPISNPAGWGDWLDFDARKTSLLMELWAANQP

ORF 13 predicted amino acid sequence

LAMFVTLKHWREFRVQRRCFIMEHKTAQVFLKVAFISCLVMSMSWLLSGNEAHAENETSV
IIQIVAGEDASIIASDHDAVVHKAIPTLGLFFVTSNHDDIQSLMANDSRVTAVYDDTIIV
GQPRFSGAVGQTLEAQPWEQGTIPSTAYSKQWATSNIRLAKAHDISQGEGTTVAILDTGI
DFDHELFQGKLVSGYDFVDNDDDPTETRDGLDQDGDLSIDEGAGHGTHVAGIIALTAPKA
NIMPIRIFDDEGRGLYFDLVAGIMYAVDNGADVINLSGSGSEDAPFLAEAVTYAEAHGVV
IVAAGAVNIYGYPASYPSVISVGASNELDYPTDFSDFPEVLNNTVYAPGFSIISSYDDGS
YAIWTGNSMATPFVAGTAALLLATNSCDDVCAKSSLLETAHHVVDDPDTSDYYGRIDAFD
AVSLATGQFHTDLNVMFMDGDSESIDDLQLKPYFNIINNGNSLPIDELTLRYWFTKDSES
EQLVECDFANVACDMIFSELGEVSETAVSDSYLELNFSPDAGILLGNHDMGDIQMRVHKS
DWTLYDEDNDYSYNGATVFTESPKITLYHNGNLVWGAEPVGAVFEPVTGEDETEDPVIIE
PPAPVVLSDVRVQYRTYDTDPSDNNIKPHFRLVNDSDTAIPLSELRVRYWFTDEATAVSQ
VHCDYAGMGCGQVTAVISATGTQHALDITFSEAAGQLGAASISGQIHTRLNHTNWQPHNE
NDDHSFLITGNDFTDWQNVTLYRNDTLIWGVEP

Appendix B 16 rRNA gene sequences of marine bacterial Isolates

Isolate 40 16S rRNA gene

tcgattagagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgagcggtaacagaaagtagcttgctactttgctgacgagcggcggacgggtgagtaat
gcttgggaacatgccttgaggtgggggacaacagttggaaacgactgctaataccgcata
atgtctacggaccaaaggggcttcggctctcgcctttagattggcccaagtgggattag
ctagttggtgaggtaatggctcaccaaggcgacgatccctagctggtttgagaggatgat
cagccacactgggactgagacacggcccagactcctacgggaggcagcagtggggaatat
tgcacaatgggcgcaagcctgatgcagccatgccgcgtgtgtgaagaaggccttcgggtt
gtaaagcactttcagtcaggaggaaaggttaacggttaatacccgttagttgtgacgtta
ctgacagaagaagcaccggctaactccgtgccagcagccgcggtaatacggagggtgcga
gcgttaatcggaattactgggcgtaaagcgtacgcaggcggtttgttaagcgagatgtga
aagccccgggctcaacctgggaactgcatttcgaactggcaaactagagtgtgatagagg
gtggtagaatttcaggtgtagcggtgaaatgcgtagagatctgaaggaataccgatggcg
aaggcagccacctgggtcaacactgacgctcatgtacgaaagcgtggggagcaaacggga
ttagataccccggtagtccacgccgtaaacgatgtctactagaagctcggagcctcggtt
ctgtttttcaaagctaacgcattaagtagaccgcctggggagtacggccgcaaggttaaa
actcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgca
acgcgaagaaccttacctacacttgacatacagagaacttaccagagatggtttggtgcc
ttcgggagctctgatacaggtgctgcatggctgtcgtcagctcgtgttgtgagatgttgg
gttaagtcccgcaacgagcgcaacccctatccttagttgctagcaggtaatgctgagaac
tctaaggagactgccggtgataaaccggaggaaggtggggacgacgtcaagtcatcatgg
cccttacgtgtagggctacacacgtgctacaatggcgcatacagagtgctgcgaacctgc
gaaggtaagcgaatcacttaaagtgcgtcgtagtccggattggagtctgcaactcgactc
catgaagtcggaatcgctagtaatcgcgtatcagaatgacgcggtgaatacgttcccggg
ccttgtacacaccgcccgtcacaccatgggagtgggttgctccagaagtagatagtctaa
ccctcgggaggacgtttaccacggagtgattcatgactggggtgaagtcgtaacaaggta
gccctaggggaacctgcggctggatcacctccttaatcactagtgaattcgcggccgcct
gcaggtcgaccata

Isolate 47 16S rRNA gene

agagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgaacggtaacatttctagcttgctagaagatgacgagtggcggacgggtgagtaatac
ttaggaatatgcctttgtgtgggggataactattggaaacgatagctaataccgcataat
gtcttcggaccaaaggggcttcggctcccgcgcaaagagtagcctaagcgagattagct
tgttggtgaggtaacggctcaccaaggcgacgatgcgtagccggcctgagagggtgaccg
gccacactgggactgagacacggcccagactcctacgggaggcagcagtagggaatcttc
cgcaatgggcgaaaccctgacggaagcgacgccgcgtgagcgaagaaggccttcgggtcg
taaagctctgttgtgagggacgaaggagcgccgttcgaagagggcggcgcggtgacggta
cctcacgaggaagccccggctaactacgtgccagcagccgcggtaatacgtaaggggcga

gcgttgtccggaattattgggcgtaaagcgcgcgcaggcggttccttaagtctgatgtga
aagcccacggctcaaccgtggagggtcattggaaactgggggacttgagtgcaggagagg
agagcggaattccacgtgtagcggtgaaatgcgtagagatgtggaggaacaccagtggcg
aaggcggctctctggcctgcaactgacgctgaggcgcgaaagcgtggggagcaaacagga
ttagataccctggtagtccacaccgtaaacgctgtctactagctgtttgtggatttaatc
cgtgagtagcgaagctaacgcgataagtagaccgcctggggagtacggccgcaaggttaa
aactcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgc
aacgcgaagaaccttacctactcttgacatactagaaacttttcagagatgaattggtgc
cttcgggaatctagatacaggtgctgcatggctgtcgtcagctcgtgtcgtgagatgttg
ggttaagtcccgcaacgagcgcaacccctgtccttagttgccagccttaagttgggcact
ctaaggagactgccggtgacaaaccggaggaaggtggggacgacgtcaagtcatcatggc
ccttacgagtagggctacacacgtgctacaatggcgagtacagagggaagcaaacttgcg
agagtaagcggatcccttaaagctcgtcgtagtccggattggagtctgcaactcgactcc
atgaagtcggaatcgctagtaatcgcaaatcagaatgttgcggtgaatacgttcccgggc
cttgtacacaccgcccgtcacaccatgggagtgggttgcaaaagaagtagctagtttaac
cttcgggaggacggttaccactttgtgattcatgactggggtgaagtcgtaacaaggtaa
ccctaggggaacctgcggctggatcacctcctt

Isolate 48 16S rRNA gene

tcgattagagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgaacggtaacatttctagcttgctagaagatgacgagtggcggacgggtgagtaatac
ttaggaatatgcctttgtgtgggggataactattggaaacgatagctaataccgcataat
gtcttcggaccaaaggggcttcggctcccgcgcaaagagtagcctaagcgagattagct
tgttggtgaggtaaaggctcaccaaggcgacgatctctagctgttctgagaggaagatca
gccacactggaactgagacacggtccagactcctacgggaggcagcagtggggaatattg
cacaatgggggaaaccctgatgcagccatgccgcgtgtgtgaagaaggccttcgggttgt
aaagcactttcagttgtgaggaaaggttaacggttaatacccgttagctgtgacgttagc
aacagaagaaggaccggctaactccgtgccagcagccgcggtaatacggagggtccgagc
gttaatcggaattactgggcgtaaagcgcacgcaggcggtttgttaagctagatgtgaaa
gccctgggctcaacctgggaattgcatttagaactggcaggctagagttttggagagggg
agtggaattccaggtgtagcggtgaaatgcgtagatatctggaggaacatcagtggcgaa
ggcgactccctggtcagtaactgacgctcatgtgcgaaagtgtgggtagcgaacaggatt
agataccctggtagtccacaccgtaaacgctgtctactagctgtttgtggatttaatccg
tgagtagcgaagctaacgcgataagtagaccgcctggggagtacggccgcaaggttaaaa
ctcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgcaa
cgcgaagaaccttacctactcttgacatactagaaacttttcagagatgaattggtgcct
tcgggaatctagatacaggtgctgcatggctgtcgtcagctcgtgtcgtgagatgttggg
ttaagtcccgcaacgagcgcaacccctgtccttagttgccagccttaagttgggcactct
aaggagactgccggtgacaaaccggaggaaggtggggacgacgtcaagtcatcatggccc
ttacgagtagggctacacacgtgctacaatggcgagtacagagggaagcaaacttgcgag
agtaagcggatcccttaaagctcgtcgtagtccggattggagtctgcaactcgactccat

gaagtcggaatcgctagtaatcgcaaatcagaatgttgcggtgaatacgttcccgggcct
tgtacacaccgcccgtcacaccatgggagtgggttgcaaaagaagtagctagtttaacct
tcgggaggacggttaccactttgtgattcatgactggggtgaagtcgtaacaaggtaacc
ctaggggaacctgcggctggatcacctccttaatcactagtgaattcgcggccgcctgca
ggtcgaccata

Isolate 53 16S rRNA gene

agagtttgatcctggctcag
gatgaacgctagcggcaggcttaacacatgcaagtcgaggggtaacagggggcttgctcc
gctgacgaccggcgcacgggtgcgtaacgcgtatacaatctgccttacactaagggatag
cccagagaaatttggattaataccttatagtttattaggatggcatcattttaataataa
agattacggtgtaagatgagtatgcgtcccattagtttgttggtaaggtaacggcttacc
aagactacgatgggtaggggccctgagagggggatcccccacactggtactgagacacgg
accagactcctacgggaggcagcagtgaggaatattggacaatgggcgagagcctgatcc
agccatgccgcgtgcaggaagacggtcctatggattgtaaactgcttttatacaggaaga
ataaggactacgtgtagtctggtgacggtactgtaagaataaggaccggctaactccgtg
ccagcagccgcggtaatacggagggtccgagcgttatccggaattattgggtttaaaggg
tccgtaggcgggctattaagtcaggggtgaaagtttgcagctcaactgtagaattgcctt
tgatactgatagtcttgaattattgtgaagtggttagaatatgtagtgtagcggtgaaat
gcatagatattacatagaataccgattgcgaaggcagatcactaacaatatattgacgct
gatggacgaaagcgtgggtagcgaacaggattagataccctggtagtccacgccgtaaac
gatggatactagctgttcggttttcggactgagcggccaagcgaaagtgataagtatccc
acctggggagtacgttcgcaagaatgaaactcaaaggaattgacgggggcccgcacaagc
ggtggagcatgtggtttaattcgatgatacgcgaggaaccttaccagggcttaaatgtag
attgacaggtttagagatagactttccttcgggcaatttacaaggtgctgcatggttgtc
gtcagctcgtgccgtgaggtgtcaggttaagtcctataacgagcgcaacccctgttgtta
gttaccagcacattatggtggggactctaacaagactgccggtgcaaaccgtgaggaagg
tggggatgacgtcaaatcatcacggcccttacgtcctgggccacacacgtgctacaatgg
taggtacagagagcagccacttagcgataaggagcgaatctataaaacctatcacagttc
ggatcggagtctgcaactcgactccgtgaagctggaatcgctagtaatcggatatcagcc
atgatccggtgaatacgttcccgggccttgtacacaccgcccgtcaagccatggaagctg
ggggtacctgaagttcgtcaccgcaaggagcgacctagggtaaaactggtaactagggct
aagtcgtaacaaggtagccgtaccggaaggtgcggctggatcacctcctt

>Isolate 54 16S rRNA gene
tcgattagagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgaacggaaacatgtctagcttgctagatgatgtcgagtggcggacgggtgagtaatac
ttaggaacatgcctttgggtgggggataactattggaaacgatagctaataccgcataac
gtctacggaccaaaggggcttcggctcccgcccagagagtggcctaagcgagattagct

agttggtgtggtaaaggctcaccaaggcgacgatctctagctgttctgagaggaagatca
gccacactggaactgagacacggtccagactcctacgggaggcagcagtggggaatattg
cacaatgggggaaaccctgatgcagccatgccgcgtgtgtgaagaaggccttcgggttgt
aaagcactttcagttgtgaggaaagtttgatggttaatacccattagatgtgacgttaac
aacagaagaaggaccggctaactccgtgccagcagccgcggtaatacggagggtccgagc
gttaatcggaattactgggcgtaaagcgcacgcaggcggtttgttaagctagatgtgaaa
gccccgggctcaacctgggaatagcatttagaactggcagactagagtcttggagagggg
agtggaatttctggtgtagcggtgaaatgcgtagatatcagaaggaacatcagtggcgaa
ggcgactccctggccaaagactgacgctcatgtgcgaaagtgtgggtagcgaacaggatt
agataccctggtagtccacaccgtaaacgctgtctactagctgtttgtggatttaatccg
tgagtagcgcagctaacgcgataagtagaccgcctggggagtacggccgcaaggttaaaa
ctcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgcaa
cgcgaagaaccttacctactcttgacatacagagaacttttcagagatgaattggtgcct
tcgggaactctgatacaggtgctgcatggctgtcgtcagctcgtgtcgtgagatgttggg
ttaagtcccgcaacgagcgcaacccttgtccttagttgccagccttaagttgggcactct
aaggagactgccggtgacaaaccggaagaaaggtggggacgacgtcaagtcatcatggcc
cttacgagtagggctacacacgtgctacaatggcgagtacagagggaagcgaacctgcga
gggtaagcggatcccttaaagctcgtcgtagtccggattggagtctgcaactcgactcca
tgaagtcggaatcgctagtaatcgcaaatcagaatgttgcggtgaatacgttcccgggcc
ttgtacacaccgcccgtcacaccatgggagtgggttgcaaaagaagtagctagtctaacc
ttcgggaggacggttaccactttgtgattcatgactggggtgaagtcgtaacaaggtaac
cctaggggaacctgcggctggatcacctcctaatcactagtgaattcgcggccgcctgca
ggtcgaccata

Isolate 56 16S rRNA gene

agagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgaacggaaacatgtctagcttgctagatgatgtcgagtggcggacgggtgagtaatac
ttaggaacatgcctttgggtgggggataactattggaaacgatagctaataccgcataac
gtctacggaccaaaggggcttcggctcccgcccagagagtggcctaagcgagattagct
agttggtgtggtaaaggctcaccaaggcgacgatctctagctgttctgagaggaagatca
gccacactggaactgagacacggtccagactcctacgggaggcagcagtggggaatattg
cacaatgggggaaaccctgatgcagccatgccgcgtgtgtgaagaaggccttcgggttgt
aaagcactttcagttgtgaggaaagtttgatggttaatacccattagatgtgacgttaac
aacagaagaaggaccggctaactccgtgccagcagccgcggtaatacggagggtccgagc
gttaatcggaattactgggcgtaaagcgcacgcaggcggtttgttaagctagatgtgaaa
gccccgggctcaacctgggaatagcatttagaactggcagactagagtcttggagagggg
agtggaatttctggtgtagcggtgaaatgcgtagatatcagaaggaacatcagtggcgaa
ggcgactccctggccaaagactgacgctcatgtgcgaaagtgtgggtagcgaacaggatt
agataccctggtagtccacaccgtaaacgctgtctactagctgtttgtggatttaatccg
tgagtagcgcagctaacgcgataagtagaccgcctggggagtacggccgcaaggttaaaa
ctcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgcaa

cgcgaagaaccttacctactcttgacatacagagaacttttcagagatgaattggtgcct
tcgggaactctgatacaggtgctgcatggctgtcgtcagctcgtgtcgtgagatgttggg
ttaagtcccgcaacgagcgcaacccttgtccttagttgccagccttaagttgggcactct
aaggagactgccggtgacaaaccggaggaaggtggggacgacgtcaagtcatcatggccc
ttacgagtagggctacacacgtgctacaatggcgagtacagagggaagcgaacctgcgag
ggtaagcggatcccttaaagctcgtcgtagtccggattggagtctgcaactcgactccat
gaagtcggaatcgctagtaatcgcaaatcagaatgttgcggtgaatacgttcccgggcct
tgtacacaccgcccgtcacaccatgggagtgggttgcaaaagaagtagctagtctaacct
tcgggaggacggttaccactttgtgattcatgactggggtgaagtcgtaacaaggtaacc
ctaggggaacctgcggctggatcacctccaa

Isolate 58 16S rRNA gene

agagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgaacggaaacatgtctagcttgctagatgatgtcgagtggcggacgggtgagtaatac
ttaggaacatgcctttgggtgggggataactattggaaacgatagctaataccgcatgac
gtctacggaccaaaggggcttcggctcccgcccagagagtggcctaagcgagattagct
agttggtgtggtaaaggctcaccaaggcgacgatctctagctgttctganaggaagatca
gccacactggaactgagacacggtccanactcctacgggaggcagcagtggggaatattg
cacaatgggggaaaccctgatgcanccatgccgcgtgtgtgaanaaggccttcgggttgt
aaagcactttcacttgtgaggaaagtttgatggttaatacccattagatgtgacgttaac
aacagaagaaggaccggctaactccgtgccagcagccgcggtaatacggagggtccgagc
gttaatcggaattactgggcgtaaagcgcacgcaggcggtttgttaagctagatgagaaa
gccccgggctcaacctgggaatagcatttagaactggcagactagagtcttggagagggg
agtggaatttctggtgtagcggtgaaatgcgtagatatcagaaggaacatcagtggcgaa
ggcgactccctggccaaagactgacgctcatgtgcgaaagtgtgggtagcgaacaggatt
agataccctggtagtccacaccgtaaacgctgtctactagctgtttgtggatttaatccg
tgagtagcgcagctaacgcgataagtagaccgcctggggagtacggccgcaaggttaaaa
ctcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgcaa
cgcgaagaaccttacctactcttgacatactagaaacttttcagagatgaattggtgcct
tcgggaatctagatacaggtgctgcatggctgtcgtcagctcgtgtcgtgagatgttggg
ttaagtcccgcaacgagcgcaacccttgtccttagttgccagccttaagttgggcactct
aaggagactgccggtgacaaaccggaggaaggtggggacgacgtcaagtcatcatggccc
ttacgagtagggctacacacgtgctacaatggcgagtacagagggaagcgaacctgcgag
ggtaagcggatcccttaaagctcgtcgtagtccggattggagtctgcaactcgactccat
gaagtcggaatcgctagtaatcgcaaatcagaatgttgcggtgaatacgttcccgggcct
tgtacacaccgcccgtcacaccatgggagtgggttgcaaaagaagtagctagtctaacct
tcgggaggacggttaccactttgtgattcatgactggggtgaagtcgtaacaaggtaacc
ctaggggaacctgcggctggtcacctcctt

Isolate 62 16S rRNA gene

agagtttgatcctggctcagattgaacgctggcggcaggcctaacacatgcaag
tcgaacggtaacatttctagcttgctagaagatgacgagtggcggacgggtgagtaatac
ttaggaatatgcctttgtgtgggggataactattggaaacgatagctaataccgcataac
gtcttcggaccaaaggggcttcggctcccgcgcaaagagtagcctaagcgagattagct
tgttggtgaggtaaaggctcaccaaggcgacgatctctagctgttctgagaggaagatca
gccacactggaactgagacacggtccagactcctacgggaggcagcagtggggaatattg
cacaatgggggaaaccctgatgcagccatgccgcgtgtgtgaagaaggccttcgggttgt
aaagcactttcagttgtgaggaaaggttaacggttaatacccgttagctgtgacgttagc
aacagaagaaggaccggctaactccgtgccagcagccgcggtaatacggagggtccgagc
gttaatcggaattactgggcgtaaagcgcacgcaggcggtttgttaagctagatgtgaaa
gccctgggctcaacctgggaattgcatttagaactggcaggctagagttttggagagggg
agtggaattccaggtgtagcggtgaaatgcgtagatatctggaggaacatcagtggcgaa
ggcgactccctggtcagtaactgacgctcatgtgcgaaagtgtgggtagcgaacaggatt
agataccctggtagtccacaccgtaaacgctgtctactagctgtttgtggatttaatccg
tgagtagcgaagctaacgcgataagtagaccgcctggggagtacggccgcaaggttaaaa
ctcaaatgaattgacgggggcccgcacaagcggtggagcatgtggtttaattcgatgcaa
cgcgaagaaccttacctactcttgacatactagaaacttttcagagatgaattggtgcct
tcgggaatctagatacaggtgctgcatggctgtcgtcagctcgtgtcgtgagatgttggg
ttaagtcccgcaacgagcgcaacccctgtccttagttgccagccttaagttgggcactct
aaggagactgccggtgacaaaccggaggaaggtggggacgacgtcaagtcatcatggccc
ttacgagtagggctacacacgtgctacaatggcgagtacagagggaagcaaacttgcgag
agtaagcggatcccttaaagctcgtcgtagtccggattggagtctgcaactcgactccat
gaagtcggaatcgctagtaatcgcaaatcagaatgttgcggtgaatacgttcccgggcct
tgtacacaccgcccgtcacaccatgggagtgggttgcaaaagaagtagctagtttaacct
tcgggaggacggttaccactttgtgattcatgactggggtgaagtcgtaacaaggtaacc
ctaggggaacctggggctggatcacctcctt

Isolate 63 16S rRNA gene

agagtttgatcctggctcag
attgaacgctggcggcaggcctaacacatgcaagtcgaacggaaacatgtctagcttgct
agatgatgtcgagtggcggacgggtgagtaatacttaggaacatgcctttgggtgggggga
taactattggaaacgatagctaataccgcatgacgtctacggaccaaaggggcttcggc
tcccgcccagagagtggcctaagcgagattagctagttggtgtggtaaaggctcaccaag
gcgacgatctctagctgttctgagaggaagatcagccacactggaactgagacacggtcc
agactcctacgggaggcagcagtggggaatattgcacaatgggggaaaccctgatgcagc
catgccgcgtgtgtgaagaaggccttcgggttgtaaagcactttcagttgtgaggaaagt
ttgatggttaatacccattagatgtgacgttaacaacagaagaaggaccggctaactccg
tgccagcagccgcggtaatacggagggtccgagcgttaatcggaattactgggcgtaaag
cgcacgcaggcggtttgttaagctagatgtgaaagccccgggctcaacctgggaatagca
tttagaactggcagactagagtcttggagaggggagtggaatttctggtgtagcggtgaa

atgcgtagatatcagaaggaacatcagtggcgaaggcgactccctggccaaagactgacg
ctcatgtgcgaaagtgtgggtagcgaacaggattagataccctggtagtccacaccgtaa
acgctgtctactagctgtttgtggatttaatccgtgagtagcgcagctaacgcgataagt
agaccgcctggagagtacggccgcaaggttaaaactcaaatgaattgacgggggcccgca
caagcggtggagcatgtggtttaattcgatgcaacgcgaagaaccttacctactcttgac
atactagaaacttttcagagatgaattggtgccttcgggaatctagatacaggtgctgca
tggctgtcgtcagctcgtgtcgtgagatgttgggttaagtcccgcaacgagcgcaaccct
tgtccttagttgccagccttaagttgggcactctaaggagactgccggtgacaaaccgga
ggaaggtggggacgacgtcaagtcatcatggcccttacgagtagggctacacacgtgcta
caatggcgagtacagagggaagcgaacctgcgagggtaagcggatcccttaaagctcgtc
gtagtccggattggagtctgcaactcgactccatgaagtcggaatcgctagtaatcgcaa
atcagaatgttgcggtgaatacgttcccgggccttgtacacaccgcccgtcacaccatgg
gagtgggttgcaaaagaagtagctagtctaaccttcgggaggacggttaccactttgtga
ttcatgactggggtgaagtcgtaacaaggtaaccctaggggaacctgcggctggatcacc
tcctt